REPORTS
OF
THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS

II

B&C SPEECH

ICA

TOKYO, AUGUST 21-28, 1968

REPORTS
of

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS

Editor in Chief
Dr. Y. KOHASI
ACOUSTICAL SOCIETY OF JAPAN

II

B & C
SPEECH



TOKYO, AUGUST 21–28, 1968

Published for the International Council of Scientific Unions with the Financial Assistance of UNESCO.

UNESCO Subvention — 1968 — DC / 2. 1 / 414 / 26

The texts of the reports contained in this book remain the property of their authors and may not be reproduced without their permission.

Printed in Japan

B

SPEECH

CONTENTS

B-1

B-1-1 Articulatory movements of the larynx
Masayuki Sawashima, Hajime Hirose, Shigeru Kiritani, Osamu Fujimura B-1

B-1-2 An electromyographic study of laryngeal activity in speech and singing
John J. Ohala, Minoru Hirano, W. Vennard ..... -5

B-1-3 What makes the vocal cords vibrate?
Kenzo Ishizaka, Masatoshi Matsudaira ..... -9

B-1-4 Perturbations in the articulatory motion of the tongue body
Robert Alphonse Houde ..... -13

B-1-5 Dynamic analysis of the vocal tract
Joseph Louis De Clerk, Jerry Salvaggi ..... -17

B-1-7 A study of dynamic palatography
Osamu Fujimura, Sadao Shibata, Zyun'ichi Simada, Shigeru Kiritani, Cotoji Satta ..... -21

B-1-8 An electromyographic study of articulatory movements
Shigeru Kiritani, Hajime Hirose, Masayuki Sawashima ..... -25

B-1-9 Temporal organization of syllable production
Björn E. F.Lindblom ..... -29

B-1-10 Sur la production des voyelles par des locuteurs hommes et femmes
René Carre, Roland Lancia, Max Wajskop ..... -31

B–2

P–0–1 Pitch extraction and display for the deaf
Ladislav Dolansky B–35

B–2–2 Automatic detection of intonational signals in American English
Yukio Takefuta ..... –39

B–2–3 Voice aperiodicity in terms of pulse shape and interval
Shinji Maeda, Osamu Fujimura, Masayuki Sawashima ..... –43

B–2–4 Fast formant frequency tracking technique
Mamoru Nakatsui, Jouji Suzuki ..... –47

B–2–5 A simple calculation of the vocal tract configuration from three formant frequencies
Yasuo Kadokawa, Jouji Suzuki ..... –51

B–2–6 On the formant information in connected speech
Yoshinari Kanamori, Shizuo Hiki, Hiroyuki Kawai, Juro Oizumi ..... –55

B–2–7 Statistical properties of speech sounds
Hiroya Fujisaki, Kunihiko Niwa, Masahiro Iwasaki, Akio Kameoka, Nobuyuki Goto ..... –59

B–2–8 Identifikation von menschlichem Verhalten durch statistische Analyse von Sprache
Fritz W. Winckel, M. Krause ..... –63

B–2–9 La segmentation temporelle : forme de prélèvement et perturbation spectrale
Albert Landercy, Max Wajskop ..... –67

B–2–10 Statistical properties of individual voices
Cesarina Bordone-Sacerdote, Gino Giacomo Sacerdote ..... –71

B–3

B–3–1 Seuils de détection de vocoïdes isolés en francais
Max Wajskop B–75

B-3-2 Harmonic components and vowel perception
Nobuo Hirato, Osamu Kakusho, Kanemasa Kato, Takashi Kobayashi B-79

B-3-3 On auditory space model of vowel perception
Hideki Kasuya, Hisayoshi Suzuki, Ken'ichi Kido ..... -83

B-3-4 Dynamic model of vowel perception
Juro Oizumi, Shizuo Hiki, Hiroshi Sato, Takeo Igarashi ..... -87

B-3-5 The role of relative spectrum levels in vowel perception
Jan Ch. K-G. Lindqvist, S. Pauli ..... -91

B-3-6 The influence of various factors on the identification and discrimination of synthetic speech sounds
Takako Kawashima, Hiroya Fujisaki ..... -95

B-3-7 Direction of transition as a perceptual parameter of time-varying stimuli
Ludmilla Andrejevna Chistovich ..... -99

B-3-8 Distinctive feature; a context-free element
Sadanand Singh .....-103

B-3-9 Is the consonant-vowel amplitude ratio an index of talker intelligibility
Arthur S. House, Michael S. Weiss .....-107

B-3-10 On the experimental evaluation of the interior redundancy in speech
Werner K. Endres .....-111

B-4

B-4-2 Several statistics of Japanese words with applications to word recognition
Shuichi Itahashi, Hisayoshi Suzuki, Ken'ichi Kido B-115

B-4-3 A spoken digit recognition system using error correcting procedure
Keiji Hiramatsu, Kaoru Kotoh B-119

B-4-4 Spoken word recognition by multiple linear separation
Seibi Chiba .....-123

B-4-5 Speech recognition with time-normalized frequency pattern
Akira Kurematsu, Seiichi Inoue .....-127

B-4-6 Discrimination of Japanese vowels in connected speech
Ken'ichi Kido, Hideki Kasuya, Hisayoshi Suzuki .....-131

B-4-7 A spoken digit recognition scheme and its computer simulation
Shigetaka Washizawa, Michio Okamoto, Takeo Shiga, Syuji Inomata 135

B-4-8 Some aspects of vowel recognition by the wave-form
Philipp Dimitrov Christov .....-139

B-4-9 Articulatory segmentation for automatic recognition of speech
Peter Bernard Denes, Theodor G. von Keller .....-143

B-4-10 Speaker identification by spectrographic voiceprints of voiceprints laboratories
Lawrence G. Kersta .....-147

B-5

B-5-1 Computer-simulated vocal organs
Eiichi Matsui B-151

B-5-2 Use of pronouncing dictionary in speech synthesis experiments
Ryunen Teranishi, Noriko Umeda .....-155

B-5-3 Synthesis of fairy tales using an analog vocal tract
Noriko Umeda, Eiichi Matsui, Torazo Suzuki, Hiroshi Omura B-159

B-5-4 A model of word and sentence intonation
Sven E. G. Öhman .....-163

B-5-5 Generalization of control rule of buzz source to improve the naturalness of synthetic speech
Takeshi Nakayama, Akira Ichikawa, Tanetoshi Miura .....-167

B-5-6 Speech synthesis by rule
Akira Ichikawa, K. Nakata .....-171

B-5-7 A hybrid control system of a human vocal tract simulator
Takashi Honda, Yasuo Ogawa, S. Inoue .....-175

B-5-8 Excitation of vocal-tract synthesizers
James Loton Flanagan, Lorinda Landgraf .....-179

B-5-9 A model for transitions of vocal tract area functions
Minoru Shigenaga, Hitoshi Ariizumi, Kiyoshi Tanaka .....-183

B-5-10 A computer controlled terminal analog speech synthesizer
Takashi Hirasaki, Hikaru Date, Hirofumi Nakajima .....-187

B-5-11 Recent experiments in parametric synthesis of polish speech sounds
Janusz Kacprowski, Wladyslaw Mikiel .....-191

B-5-12 Speech synthesis system based on interphoneme transition unit
Shin'ichiro Hashimoto, Shuzo Saito .....-195

B-5-13 Speech synthesis by rule supplementarily using natural speech segments
Yasuo Kato, Kazuo Ochiai, Shintaro Azami .....-199

B-5-14 New results in speech synthesis by rule
Lawrence Richard Rabiner, Harry Levitt B-203

B-5-15 Speech synthesis using a structured phoneme classification scheme
P. V. S. Rao, R. B. Thosar .....-207

B-5-16 A multiplexing speech synthesizer
Kazuo Nakata, Tanetoshi Miura .....-211

B-5-17 The use of teaching machines for training experiments in speech perception
Karl-Erik Spens, Arne Risberg .....-215

B-5-18 Periodic-nonperiodic test of hearing capacity
Arne Risberg .....-219

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-1-1

Articulatory Movements of the Larynx

Masayuki Sawashima, Hajime Hirose,
Shigeru Kiritani and Osamu Fujimura
Research Institute of Logopedics and Phoniatrics,
Faculty of Medicine, University of Tokyo

The larynx in articulation of Japanese consonants has been observed by use of a flexible fiberscope inserted through the nose. Photo-electric glottograms has been obtained simultaneously with cinematography and sound recordings.

Experimental Techniques

The fiberscope consists of a hard tip, a flexible optical cable, an optical connector, a light source and a camera (Fig. 1). The cable with its hard tip is approximately 6-mm in diameter. The opticale contains a coherent fiber-optics as the image guide, and a fiber-optic light guide.

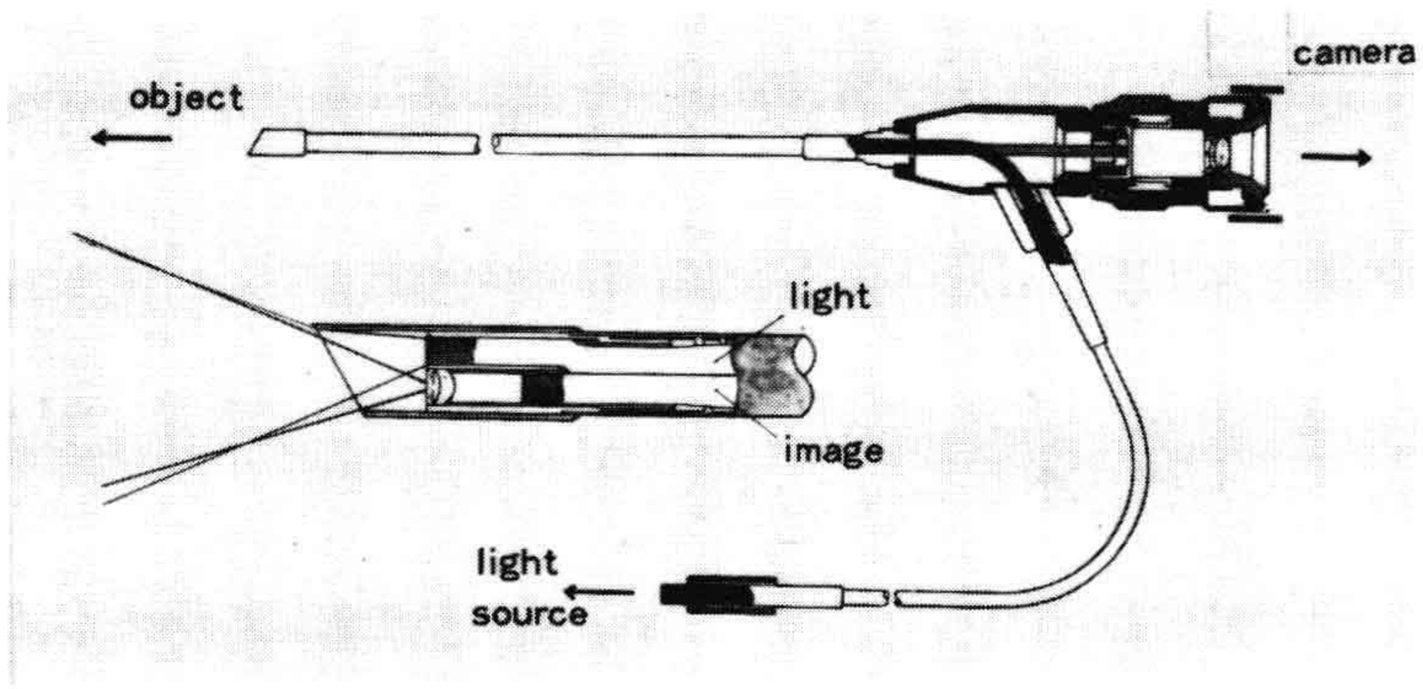


Fig. 1 Structure of the fiberscope

Before the insertion of the scope, the subject is anesthetized in the nasal fossa and the epipharynx with 4% Xylocain with Epinephrin. Unless the nasal cavity of the subject is too narrow, the insertion of the optical cable is carried out with ease and causes no discomfort or disturbance to the subject in usual pronunciation. When the tip of the scope reaches near the level of the tip of the epiglottis, the larynx is readily visible through the image finder of the camera. The scope can be kept in position and repositioned by manipulating the optical cable at the nostril. Fig. 2 shows still pictures of the glottis of a normal adult male subject in a stationary pronunciation of [s_J], and of [i] in falsetto.

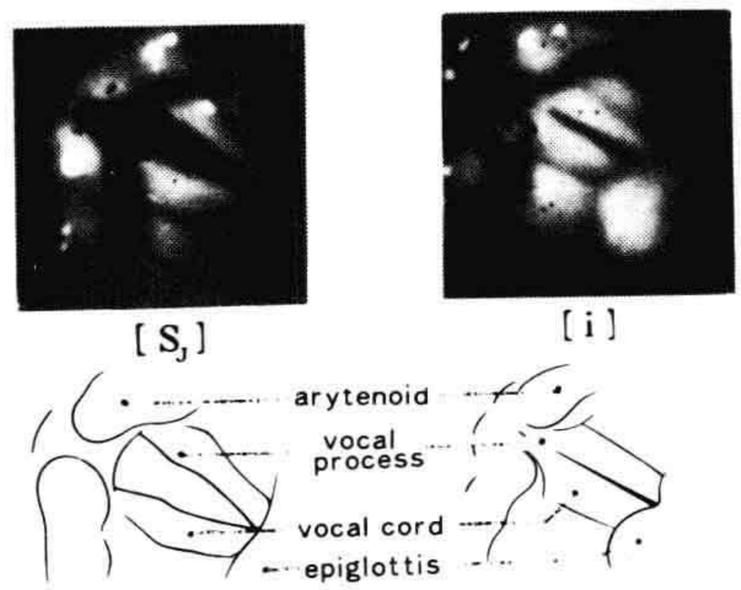


Fig. 2 The glottis viewed through the fiberscope

In the present study, a 150-watt incandecent lamp with a 20 v DC supply is used as the light source. A motion picture with synchronizing time marks is taken with a frame rate of 24 per second. During the process of cinematography, a photo-multiplier tube is placed against the pre-tracheal wall of the subject just bellow the larynx. The light conducted through the light guide passes through the larynx and transcutaneously reaches the photo-multiplier tube. The output signals of the photo-tube are recorded in different levels in three channels of 7-channel FM tape recorder. The frequency response of the FM tape recorder is flatt up to $\pm$1 dB from DC to 5000-Hz. The AC component of the photo-electric signal is recorded also by a dual-track tape recorder simultaneously with the sound and the timing marks. Fig. 3 is a block diagram which outlines the arrangement of the equipments.

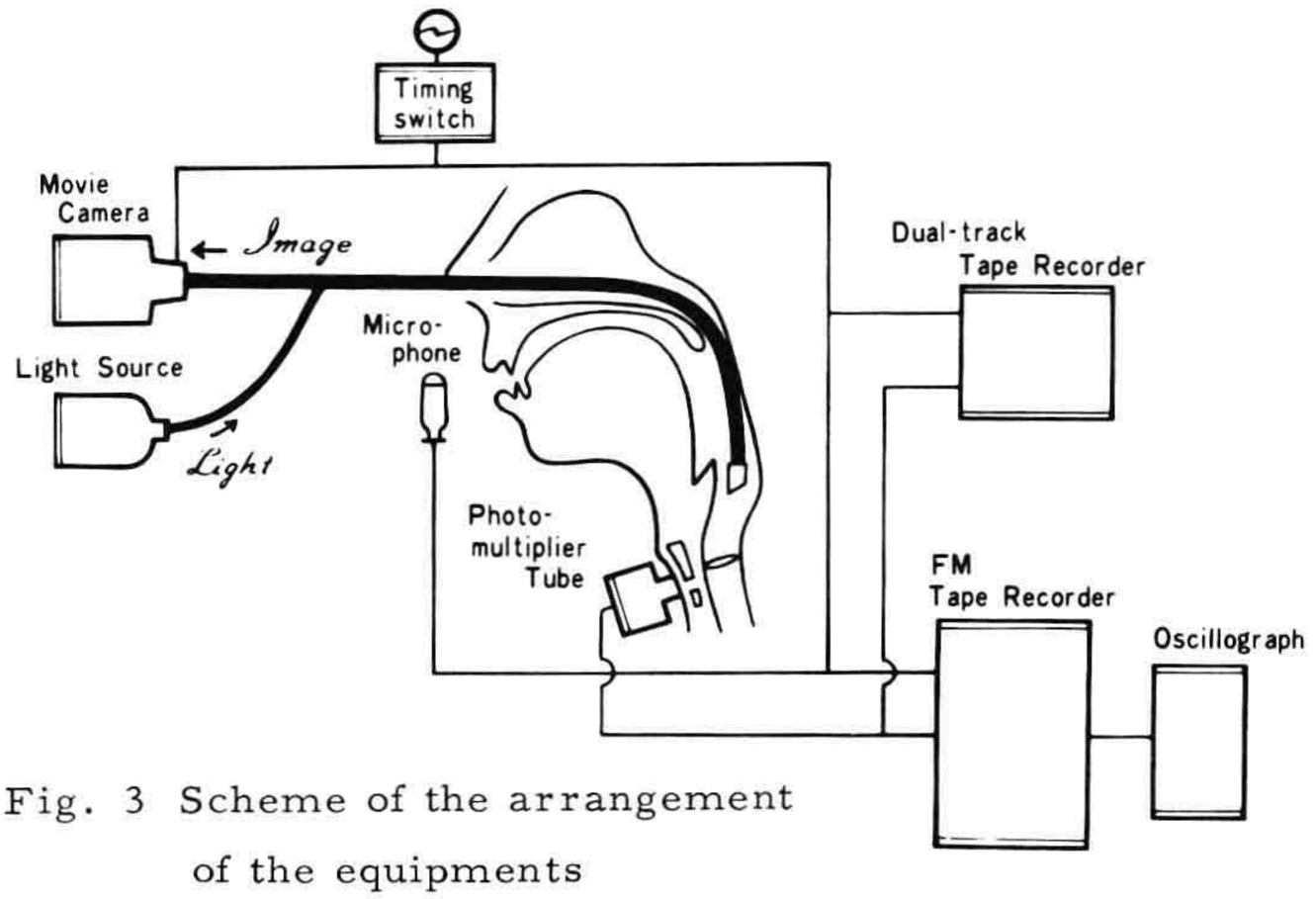


Fig. 3 Scheme of the arrangement of the equipments

When the glottis is open, a linear relationship is observed between the output level of the photo-electric signal and the area of the glottis obtained from the motion picture. The output level shifts with the variation in positioning of the optical cable which contains the light guide. Visual or photographic monitoring of the glottis is indispensable for the correct interpretation of the photo-electric glottogram.

Results

In this study of consonant articulations, CVCV syllables are embedded in the carrier sentence as "Kore wa ii CVCV desu." The symbol C in this case may represent a lax, tense or a geminate tense consonant. Fig. 4 is a recorded sample of a /keɑke/, where /ɑk/ is manifested phonetically as a geminate consonant [kk]. The subject is a normal Japanese male adult. A sound spectrogram of the speech sound, sampled frames of the motion picture and a photo-electric glottogram are shown. An upward shift of the curve in the glottogram represents an opening process of the glottis, and downward shift, a closing process.

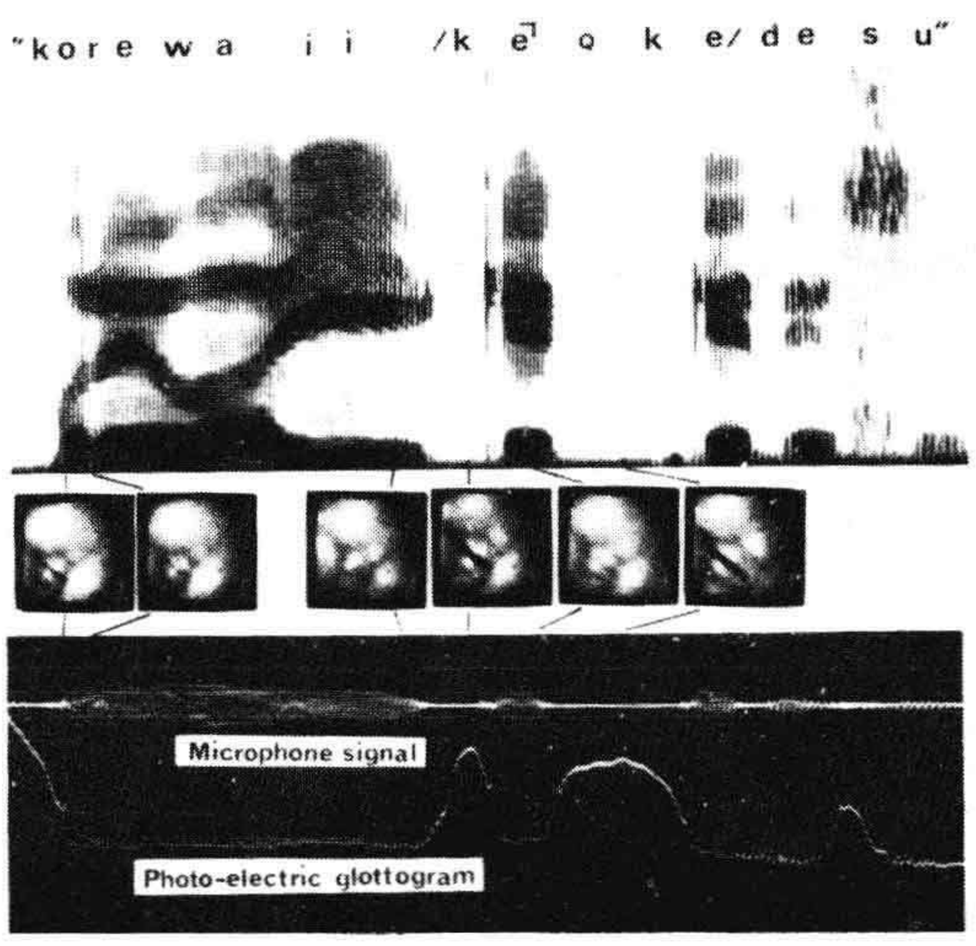


Fig. 4 Recorded sample of an utterance

In the constrictive phases of voiceless consonants, the vocal cords are kept apart, and the size of the maximum glottal area generally seems to be greater in fricative consonants than in stops in the same phonological environment. The maximum area also seems to vary in accordance with the phonetic environment.

Vibrations of the vocal cords cease gradually with the opening movement of the vocal cords after a closure or constriction of the vocal tract has been initiated. This point is more readily observable in the second syllable of CVCV.

An abrupt buildup of the vocal cord vibration is observed for a vowel portion following a consonant.

In [h], the vocal cords are apt to vibrate throughout the consonantal period even while they are clearly apart.

In the voiced stops and fricatives, compared to vowel segments, a reduction of the amplitude of the vocal cord vibration is observed. No marked change in the glottal condition is observable for nasal consonants.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-1-2

An Electromyographic Study of Laryngeal Activity in Speech and Singing

John Ohala
Research Assistant, Department of Linguistics
University of California, Los Angeles
Minoru Hirano
Assistant Research Phoniatrician
University of California, Los Angeles
William Vennard
Chairman, Department of Voice, University of Southern Calif.

An understanding of exactly how the fundamental frequency (Fo) and intensity of voice are controlled during speaking and singing is important for the laryngologist, the linguist, the language teacher, and the singing teacher.

Towards this end we studied the activity of five main intrinsic laryngeal muscles and selected extrinsic laryngeal muscles. Electromyographic recordings were obtained from these muscles during singing and running speech--painlessly--by use of bipolar thin wire electrodes similar to those described by Basmajian and Stecko (1962). The method of electrode insertion into the intrinsic laryngeal muscles was that of the transcutaneous approach developed by Hiroto, Hirano, Toyozumi, and Shin (1962). Techniques for transcutaneous electrode insertion into the extrinsic laryngeal muscles were developed after extensive anatomical study on human cadavers. Processing of the electromyographic data obtained during speech was aided by a LINC-8 computer which rectified, superimposed and smoothed the records from several tokens of the same utterance. (fig.3) Five speakers of English, four males and one female, served as subjects although not all of the above muscles

were investigated in each subject. Two of these subjects, one male and one female, were trained professional singers.

Basically our results led us to classify the intrinsic laryngeal muscles into three categories on the basis of function: the adductor muscles (interarytenoid, lateral cricoarytenoid and vocalis muscles), the abductor muscle (posterior cricoarytenoid) and the tensor muscle (cricothyroid muscle). This conforms to the traditional classification of the intrinsic laryngeal muscles. However, the situation is more complicated than this. The participation of individual muscles in various gestures in speech and singing can not always be easily analyzed simply and only as adduction, abduction or tensing. For example, the vocalis muscle figures in the regulation of registers of voice as well as in purely adductory gestures such as glottal stops. Further, the lateral cricoarytenoid muscle is active, as expected, during voicing but also shows increasing activity for increasing Fo (Fig.3)

Simple gestures in speech such as opening and closing the glottis reveal themselves as complex, precise actions involving several muscles. To move the vocal cords from one position to another requires not merely that a given muscle contract, but also that its antagonist be inhibited. In Fig. 1, is shown the electromyographic activity of the pair of antagonistic muscles, the lateral and the posterior cricoarytenoid muscles, during the production of a series of glottal stops. The reciprocal activation of these two muscles is strikingly precise. In Fig. 2, the activity of the interarytenoid muscle is shown during production of the word 'behave'. General activation is present for all voiced portions of the word, but a clearly defined inhibition occurs for the production of the 'h'. That such precise synchronized on-off activity of many muscles is also capable of being exactly repeated

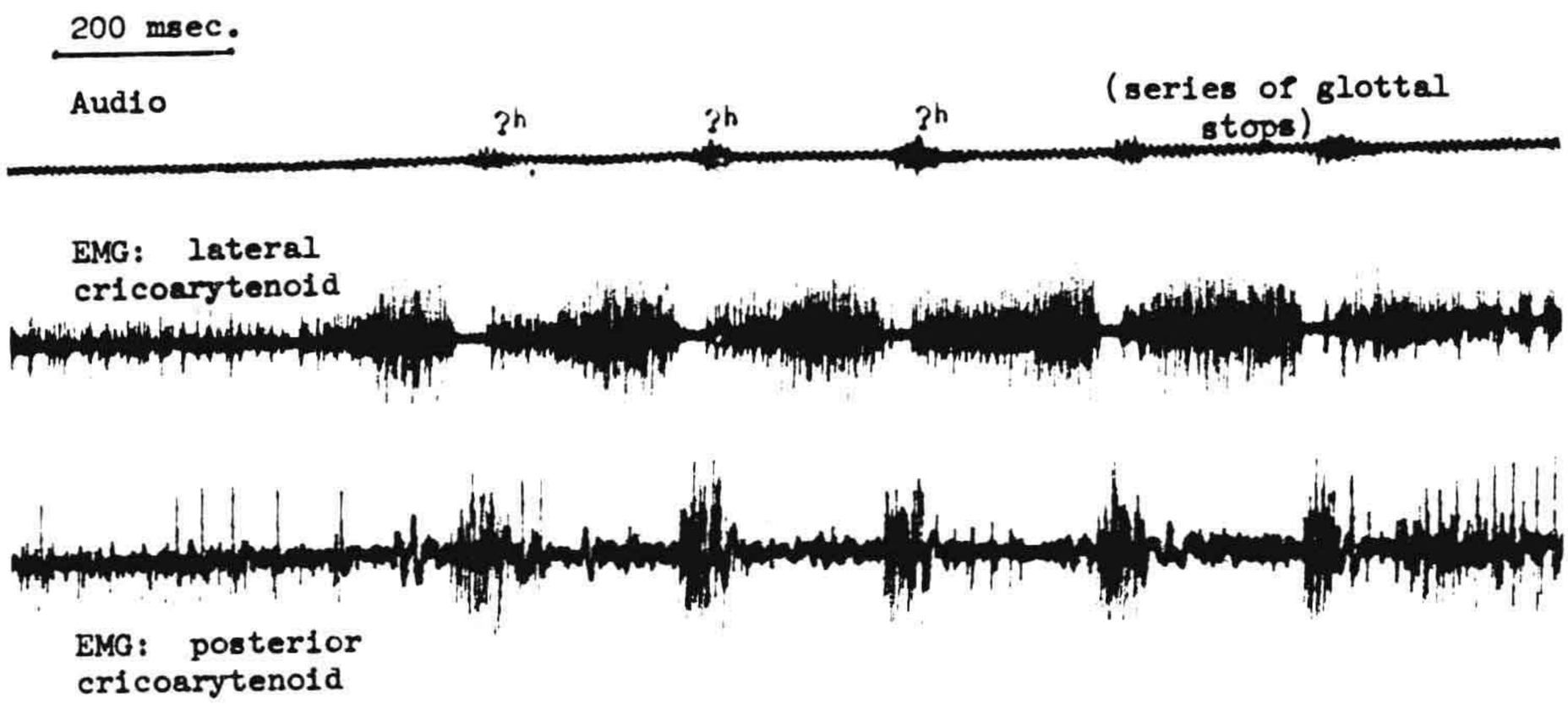


Fig. 1

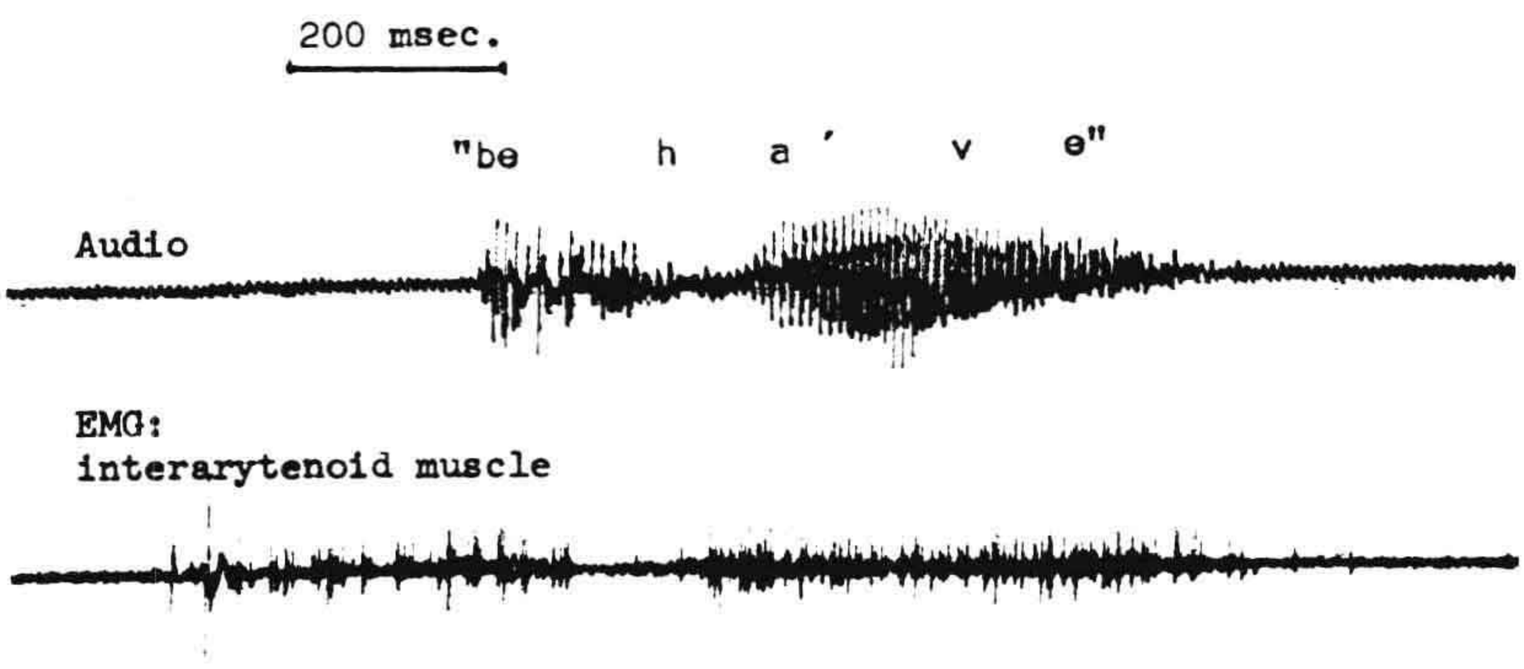


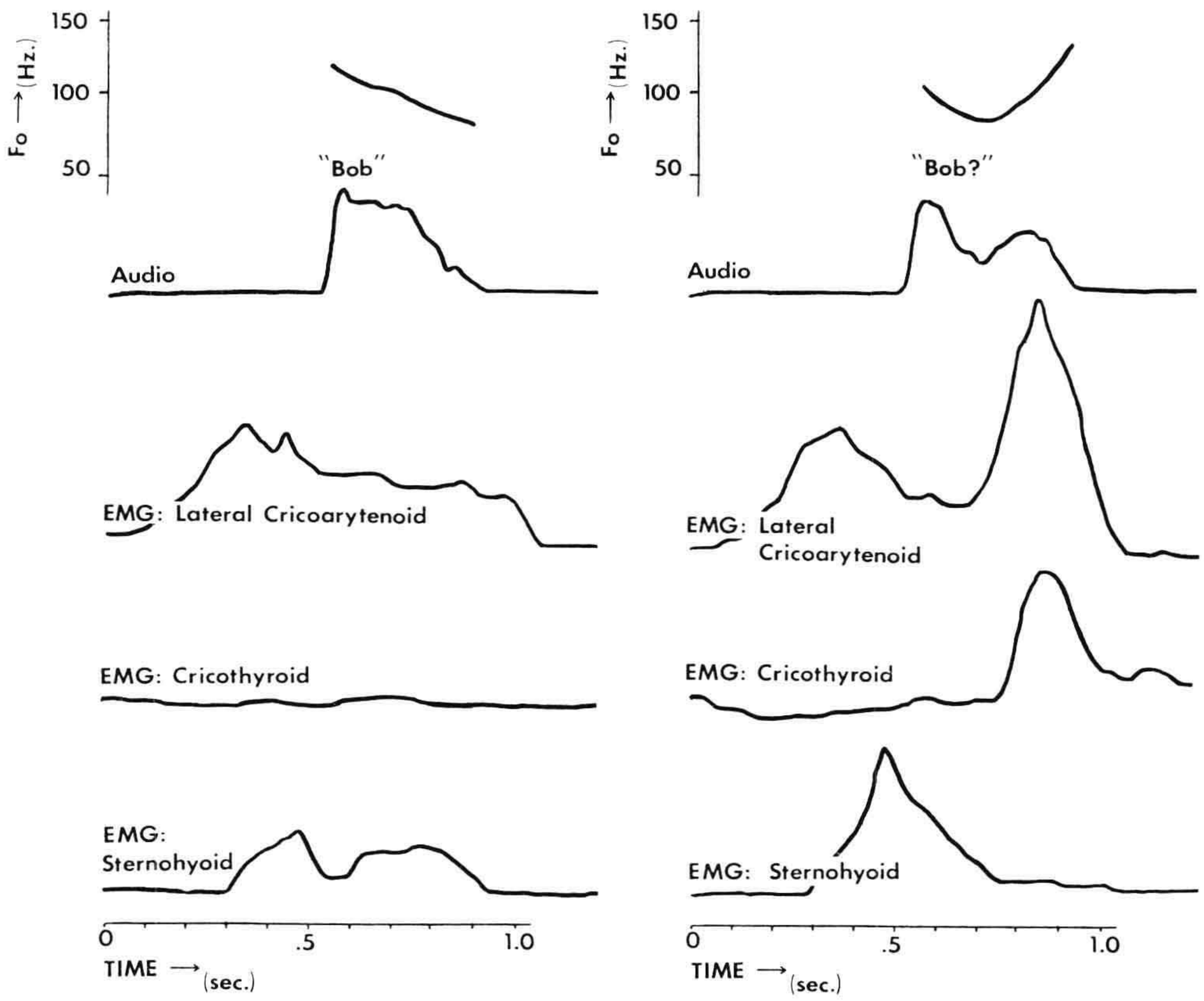
Fig.2

time after time is drāmatic visual evidence that speech is one of man's most highly skilled activities and furthermore argues for the existence of an underlying stable articulatory 'program' which controls these gestures each time a given word is uttered.

The chief muscle regulating Fo is the cricothyroid muscle, but all of the adductory muscles also participate. As far as the laryngeal muscles are concerned intensity is chiefly regulated by the adductor muscles, especially the vocalis muscle. Variations in the pattern of Fo and intensity regulation appear in the different registers and fundamental frequencies. For example, in extremely high-pitched falsetto, participation of the laryngeal muscle seems to be less important for varying Fo and intensity; thus the air flow

rate must be the more dominant means of control in this case. Other than this the control mechanisms for Fo and intensity did not differ for normal speech and for singing in the chest register.

The pattern of activity of the extrinsic laryngeal muscles in speech is more difficult to determine due to their apparent simultaneous participation in more than one gesture. One clear pattern which did appear is the activation of the sternohyoid muscles in the lowering of Fo (Fig. 3).



References

Basmajian,J.V.,and Stecko,G.(1962)A new bipolar electrode for electromyography,J.Appl.Physiol.17.849

Hiroto,I.,Hirano,M.,Toyozumi,Y.,and Shin,K.(1962) A new method of placement of a needle electrode in the intrinsic laryngeal muscles for electromyography,Oto-rhino-laryngol.Clinic kyoto 55,499-504 (in Japanese)

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-1-3

What makes the Vocal Cords vibrate

Kenzo Ishizaka, University of Electro-Communications,
Masatoshi Matsudaira, Faculty of Engineering, Tamagawa University.

In earlier work[1] the mechanism of self-excited vibration of the larynx in the chest voice has been studied by a single-degree-of-freedom model, as shown in Fig. 1. Measurements of the voice pitch for different acoustic loads, however, have shown that the conditions for oscillation of the single-degree-of-freedom model do not apply to vibration of the vocal cords.[2]

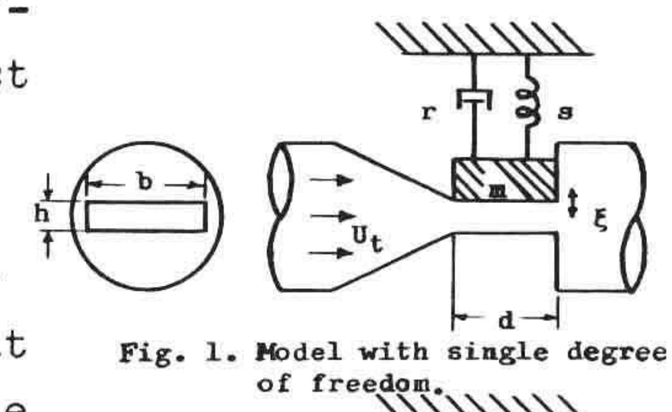


Fig. 1. Model with single degree of freedom.

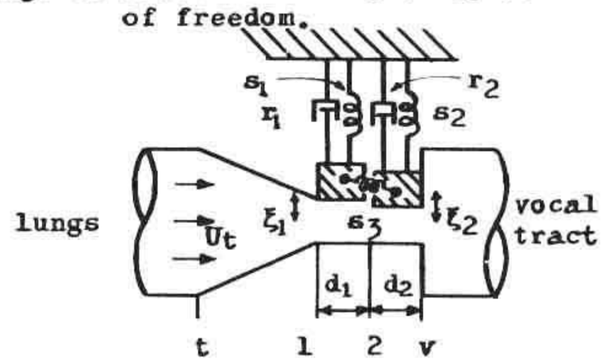


Fig. 2. Model with two degrees of freedom.

In this report a two-degree-of-freedom model is proposed, as shown in Fig. 2, and it is given that this idealized model has several characteristics of oscillation in common with the vocal cords. It is assumed that masses m_1 and m_2 are capable of purely vertical motion in Fig. 2, and so the coupling stiffness s_3 acts to supply a force on m_1 and m_2 only in the vertical direction by virtue of a difference in their displacements ξ_1 and ξ_2, respectively. To investigate small oscillatory components of the air flow and the pressure caused by the small vibrations of m_1 and m_2 superimposed on a parallel glottal slit, we assume that the flow in the glottis is two-dimensional parallel flow between the vibrating walls, and that the velocity profile of the steady flow is uniform across the section, i.e., that the mean velocity is U , where U is much smaller than the velocity of sound. Furthermore, we assume that the flow through the contraction or the subglottis is one-dimensional lossless flow between the vibrating walls. The approximate small-signal equiva-

lent circuit for the excess pressure, p, and the mean superimposed volume velocity, q, is constructed from the solutions of the Navier-Stokes equation and the equation of continuity, as is shown in Fig.3, which is simplified by the omission of parallel elements. The acoustic driving point impedances looking into the trachea and vocal tract are denoted by Z_t and Z_v, respectively.

The equivalent pressure sources $E_3/2$ are inserted to fulfil the conditions at the boundary between m_1 and m_2. The net excess pressures F_1 and F_2 acting on m_1 and m_2, respectively, are given by the pressures p_1, p_2, and p_v, where F_1 and F_2 are taken to have twice the values due to only m_1 and m_2 because of the symmetrical vibration of both of the vocal cords. The force equations of the mechanical system of the model are as follows;

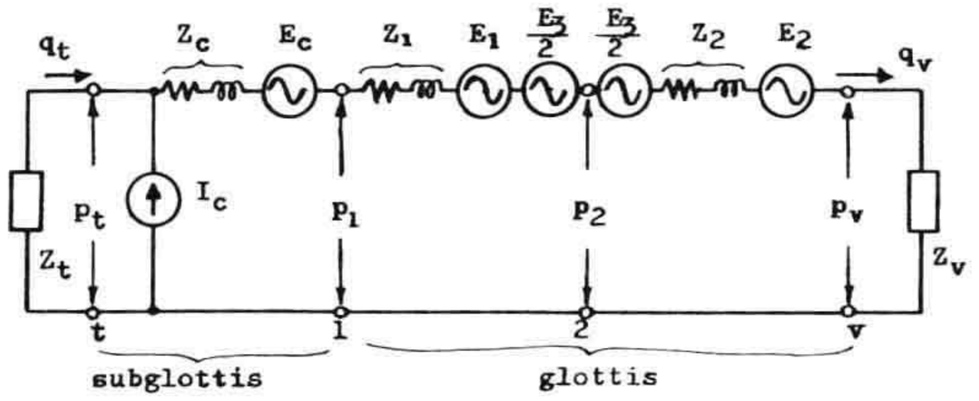


Fig. 3. Simplified small-signal equivalent circuit.

$$Z_{m_1}\dot{\xi}_1 + s_3(\xi_1-\xi_2) = F_1 \approx \frac{A}{2}\left[3(Z_t+Z_g+Z_v)\xi_1+\left\{Z_v+ 4N\cdot R_c-3(Z_t+Z_g)\right\}\xi_2\right]$$
$$Z_{m_2}\dot{\xi}_2 + s_3(\xi_2-\xi_1) = F_2 \approx \frac{A}{2}\left[\left\{3Z_v+ 4N\cdot R_c-(Z_t+Z_g)\right\}\xi_2+(Z_t+Z_g+Z_v)\xi_1\right] \qquad (1)$$

where $Z_{m_1} = r_1 + j(\omega m_1 - s_1/\omega)$, $Z_{m2} = r_2 + j(\omega m_2 - s_2/\omega)$,

$Z_g = Z_c + Z_1 + Z_2$ is the total glottal slit impedance, $A = b^2dU/2$, and $N = S/S_t$, the ratio of the cross-sectional area of the glottis to that of the trachea. Since $R_c = \rho U/S \gtrsim 40$ cgs acoustic ohms and $Z_g \approx R_c \gg |Z_t|$, $|Z_v|$ for the fundamental frequency of the vocal cord vibration, Z_t and Z_v will be disregarded here for simplicity. The case is different in the single-degree-of-freedom model, whose force equation can be reduced from Eq.(1) as a special case of $s_3 = \infty$, i.e., $\xi_1 = \xi_2$, as follows;

$$Z_{m_1} + Z_{m2} = \frac{A}{j\omega}(4Z_v + 4N\cdot R_c) \qquad (2)$$

In this case Z_v is predominant in the oscillation condition.

The characteristic equation of Eq.(1) is given by

$$\lambda^4 + 2(\beta_1+\beta_2)\lambda^3 + (\omega_{o1}^2+\omega_{o2}^2+4\beta_1\beta_2)\lambda^2 + 2(\beta_1\omega_{o2}^2+\beta_2\omega_{o1}^2)\lambda + \omega_{o1}^2\omega_{o2}^2(1 \pm k^2) = 0 \qquad (3)$$

where $\beta_1 = \frac{r_1}{2m_1}$, $\omega_{o1}^2 = \frac{s_1+s_3-\frac{3}{2}s_4}{m_1}$, $s_4 = A\cdot R_c = (bd)\cdot\rho U^2/2h$,

$\beta_2 = \frac{r_2}{2m_2}$, $\omega_{o2}^2 = \frac{s_2+s_3+\frac{1}{2}s_4}{m_2}$, $\pm k^2 = \frac{(\frac{3}{2}s_4-s_3)(\frac{1}{2}s_4+s_3)}{(s_1+s_3-\frac{3}{2}s_4)(s_2+s_3+\frac{1}{2}s_4)}$,

and k is the coupling coefficient of the system and is positive or negative depending on the value of s_4 , whose numerical example is shown in Fig.4 . The instability condition of Eq.(3) requires that

$$(\omega_{o1}^2 - \omega_{o2}^2)^2 + 4(\beta_1+\beta_2)(\beta_1\omega_{o2}^2+\beta_2\omega_{o1}^2) - \frac{(\beta_1+\beta_2)^2}{\beta_1\beta_2}\omega_{o1}^2\omega_{o2}^2 < 0 \quad . \qquad (4)$$

This necessarily means that $(3/2)s_4 > s_3$ and that the two frequencies ω_{o1} and ω_{o2} are equal or nearly equal.

In the simple case of $\beta_1 = \beta_2 = \beta$ and $\omega_{o1} = \omega_{o2} = \omega$, if

$$\frac{k\omega}{2} - \beta > 0 \quad \text{or} \quad k > 2\zeta \quad , \qquad (5)$$

then a pair of roots of Eq.(3) have a positive real part as follows,

$$\left.\begin{matrix}\lambda_1 \\ \lambda_2\end{matrix}\right\} = \left(\frac{k\omega}{2} - \beta\right) \pm j\omega\sqrt{1 + k^2/4} \qquad (6)$$

in which $\zeta = \beta/\omega = 1/2Q$ denotes the damping ratio. Eq.(5) is the condition for self-excited oscillations.

From Eq.(1) and (6), the ratio between $\dot{\xi}_1$ and $\dot{\xi}_2$ becomes

$$\dot{\xi}_1 / \dot{\xi}_2 = j\left(1 + \frac{s_2}{s_3 + \frac{1}{2}s_4}\right)k\sqrt{1 + k^2/4} \qquad (7)$$

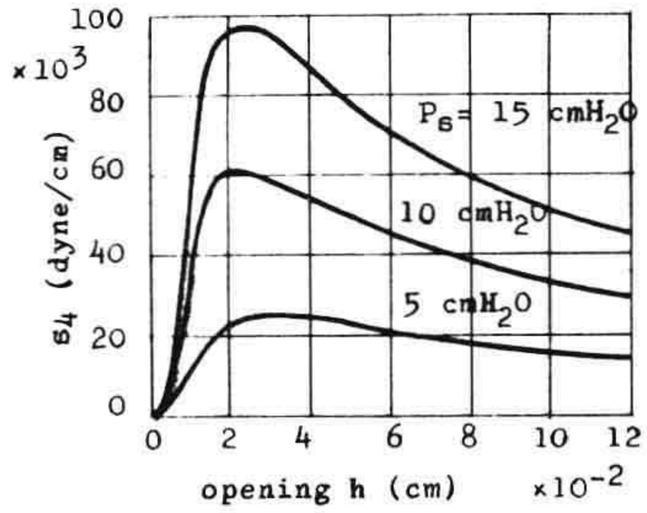


Fig. 4. Apparent stiffness s_4

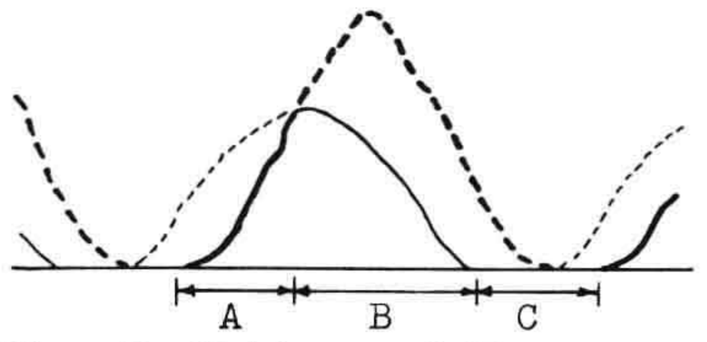


Fig. 5. Motions of the upper (heavy curve) and lower(light curve) margins of the vocal cords (after Hiroto, 1966)

This result shows that the velocity $\dot{\xi}_1$ advances with respect to the velocity $\dot{\xi}_2$ by phase angle 90°, which practically corresponds to the observed phase difference in the motion between the upper and lower edges of the vocal cords, as is shown in Fig. 5, and which has been inferred by Hiroto from his high speed motion pictures. [4]

To satisfy the condition $\omega_{o1} \approx \omega_{o2}$ in the actual larynx, it is necessary that the lower margin of the vocal cords effectively be much more tense than the upper margin, which is probable in practice because of the looseness of the upper surface of the vocal cords. Then the condition may be fulfilled by the favorable adjustment of s_4, i.e., of the glottal opening and the subglottal pressure.

From measuring data on the static tension of the upper and lower margins of the vocal cords, the values of their equivalent stiffnesses are estimated to be

$s_1 = 60$ k·dyne/cm, $s_2 = 24$ k·dyne/cm, $s_3 = 12$ k·dyne/cm

for the tensions of the mucous membrances only, therefore these values must be applied to the lower frequency of the vocal cord vibration. [5]

The calculated logarithmic increment of the model for these values of s_1 , s_2 , and s_3 are tabulated in Table 1,
for $\omega_{o1} = \omega_{o2} = 2\pi(120\ \text{Hz})$, $m_1 = m_2 = 0.08$ g, $h = 0.05$ cm

Table 1.

		two-degree-of-freedom model		single-degree-of-freedom model
velocity of steady flow U (m/s)	subglottal pressure P_s (cm H_2O)	coupling coefficient k	logarithmic increment for $\zeta=0.1$ $2\pi(k/2-\zeta)$	logarithmic increment for $\zeta=0$ $2\pi(4AL_v/2mw)$
15	2.9	0.168	$2\pi(-0.016)=-0.10$	$2\pi(0.032)=0.20$
20	4.3	0.42	$2\pi(0.11)= 0.7$	$2\pi(0.043)=0.27$
30	8.6	1.83	$2\pi(0.60)= 3.8$	$2\pi(0.063)=0.40$

For U = 20 m/sec, the amplitudes of the oscillations build up by a factor of about two during one period, and when U is more than 30 m/s, the harmonic oscillations rapidly tend toward relaxation oscillations.

The increase of subglottal pressure from 4.3 cm H_2O to 8.6 cm H_2O produces an increase in the oscillation frequency by about 33 percent, which is comparable with the experimental data of the excised larynx.

For comparison, the table also gives a numerical example of the logarithmic increment for $\zeta = 0$ of the single-degree-of-freedom model with a uniform pipe, 17 cm long and 4 cm^2 cross-sectional area. In this case, if the damping ratio ζ is of the order of 0.1 , the subglottal pressure is required to be above 20 cm H_2O for the expanding oscillation, or self-excited oscillations can exist at the normal subglottal pressure only when the vocal cords have a considerably higher Q than is reasonable. Moreover, the oscillations cannot be self-excited for Z_v of capacitive reactance, according to Eq.(2). This fact is confirmed by the experiments with mechanical models constructed similar to the single-degree-of-freedom model, which are driven by air flows, but is not true for the actual larynx, where measurements have shown that we could sustain the vowel sound having no bearing on the character of the reactance of Z_v .[3]

The two-degree-of-freedom model similarly has little effect of the acoustic load on the condition for sustained oscillation, and has several characteristics of the vibration corresponding to those of the larynx. It therefore appears that the mechanism of the self-excited oscillation represented here by this idealized model reveals the intrinsic one of the larynx in the chest voice: the phase difference of the margins is not a parasistic phenomena caused by the opening and closing motion of the vocal cords, but has a significant effect by means of which the larynx can be easily and strongly self-excited.

Thanks are due to Prof. O. Fujimura for his interest and help,and to Dr. T. Kaneko for supplying the data concerning the vocal cords.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-1-4

Perturbations in the Articulatory Motion of the Tongue Body

Robert A. Houde

Research Staff, General Dynamics Electronics Division

Rochester, New York, U.S.A.

Introduction

This report describes some results of a study of tongue body motion.[1] The purpose of this study was to observe tongue body behavior during speech production and to relate this behavior to the phonemic representation of the corresponding speech signals. The tongue body was selected for study because of its importance in the speech process, because of its relative simplicity and particularly because of its amenability to the marking technique employed. Radiopaque markers were attached to the surface of the tongue body in order to identify particular surface points through lateral cineradiography. The tongue body was represented in terms of surface points rather than in terms of its shape or the related area function in anticipation that the point motions would more directly reflect the mechanico-inertial constraints of the tongue and result in a simpler characterization. The orientation of the study was toward an analysis in terms of articulatory phoneme targets. In addition to the principal component of motion which could be related to phoneme targets, a component of motion more directly related to the state characteristics of the vocal mechanism such as pressure, tongue configuration and position was also observed.

Experimental Technique

Six points on the articulators of a single subject were marked by the attachment of small radiopaque markers. The motion of these markers during the production of a set of utterances was recorded by lateral cineradiography at 100 frames per second. The locations of the six markers are shown in Figure 1. Four are located on the body of the tongue and one is located on each lip. In addition, a distinct dental filling was designated as a marker for the lower jaw. The utterances employed were the 18 symmetrical trisyllables which can be constructed from the phonemes /i/, /ɑ/, /u/, /b/, /g/ in the form VCVCV. Each of these 18 utterances was produced with stress on the initial, middle, and final syllables. The resulting film was reduced by frame-by-frame analysis to point marker locations which were plotted as horizontal and vertical displacements vs. time and also as paths in the mid-sagittal plane.

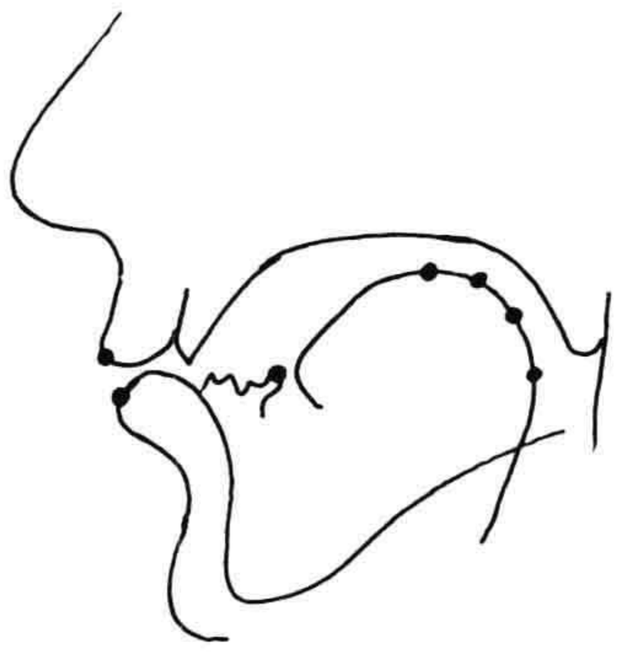


Figure 1. Point marker locations.

Data Analysis

The general approach to analysis was (1) For each tongue point in each utterance, target positions were identified corresponding to each phoneme except /b/. Vowel target position was defined as that position sufficiently sustained. (less than 1/2 mm change in 100 ms). In the case of /g/ the target was considered to be that position initially taken during palatal contact. (2) The motion of a tongue point along the axis connecting consecutive target positions was identified as the target-directed component of motion. (3) The motion of the tongue point along the axis perpendicular to the inter-target axis was identified as the deviation component of motion. It was found that the target directed component of tongue point motion for those cases involving consecutive vowel targets could be represented by a single

characteristic transition function independent of the direction of the motion or the timing of consecutive vowels. A second characteristic transition function was found to be sufficient to represent target-directed motion in those cases involving /g/. Detailed analysis of the target-directed components was reported previously.

The deviation components during the VbVbV utterances were observed to be generally in the form of short "perturbations" synchronized with labial closure. The size of the perturbations varied from a minimum when the deviation component axis was horizontal, to a maximum downward when the axis was vertical, thus suggesting a downward directed perturbation. These perturbations were isolated from the total vertical motion of each tongue point by subtracting from the total motion, the target-directed motion predicted by the characteristic transition function. Figure 2 shows some of the perturbation components isolated in this way for the frontmost tongue point. The tongue body is observed to move downward a few mm during labial closure and then to return to its initial position following the release of closure. The perturbations observed in the motion of the other three tongue points were similar in form and timing which indicates that the tongue body is perturbed as unit. The possibility that these labial closure-related perturbations were due to motion of the jaw or of the laryngeal structure was considered and rejected. Poor correlation was observed between the motion of these structures and tongue body perturbations. The dependence of the perturbations on oral pressure variations was established experimentally by observations during production of /b/ with and without a pressure leak tube between the lips. A similar analysis was carried out for those utterances involving palatal closures. In this case the perturbation was observed to take place horizontally during closure in the forward direction. Figure 3 shows some perturbations isolated from utterances involving /g/.

Conclusions

Oral pressure perturbations produce significant perturbations in tongue body motion. During bilabial closure the tongue is displaced downward and during palatal closure the tongue is displaced forward. The size of these displacements is sufficient to account for the compliance in the oral cavity required to sustain voicing during voiced stops. The effect on the acoustic signal is small in the case of labial closure. A displacement of a few mm in tongue height can, in some cases, produce formant frequency shifts equal to their difference limen. In the case of palatal closures the perturbations are generally larger and have a greater effect on the acoustic signal. A relationship between the size of the perturbation and tongue position is indicated, however the evidence is not conclusive.

1. R. A. Houde, "A Study of Tongue Body Motion During Selected Speech Sounds," Ph. D. dissertation, University of Michigan, 1967.

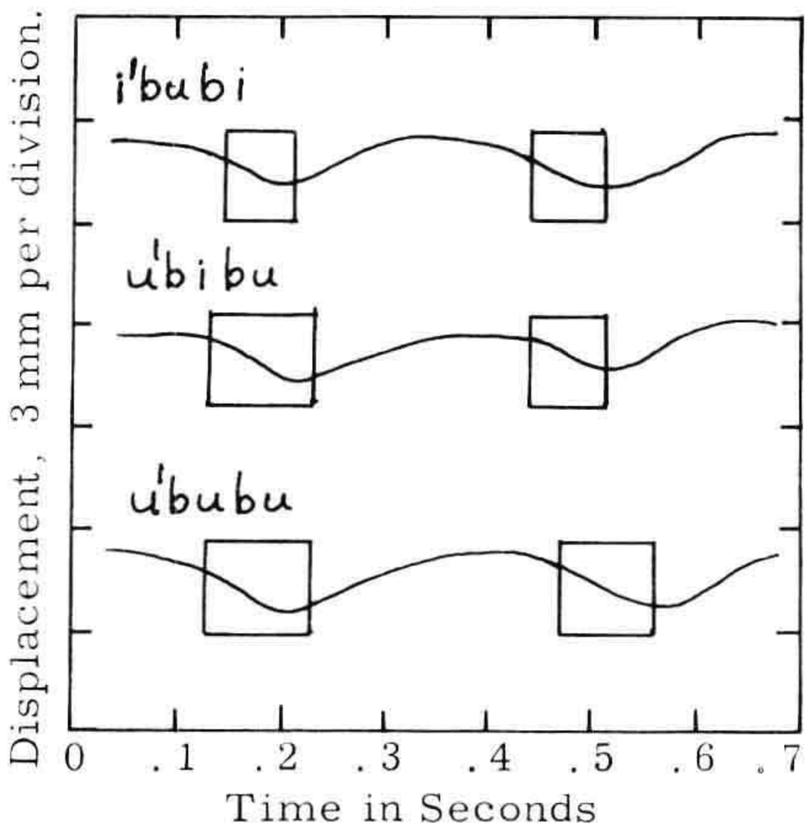


Figure. 2. Downward directed perturbations during VbVbV utterances. Boxes indicate periods of labial closure.

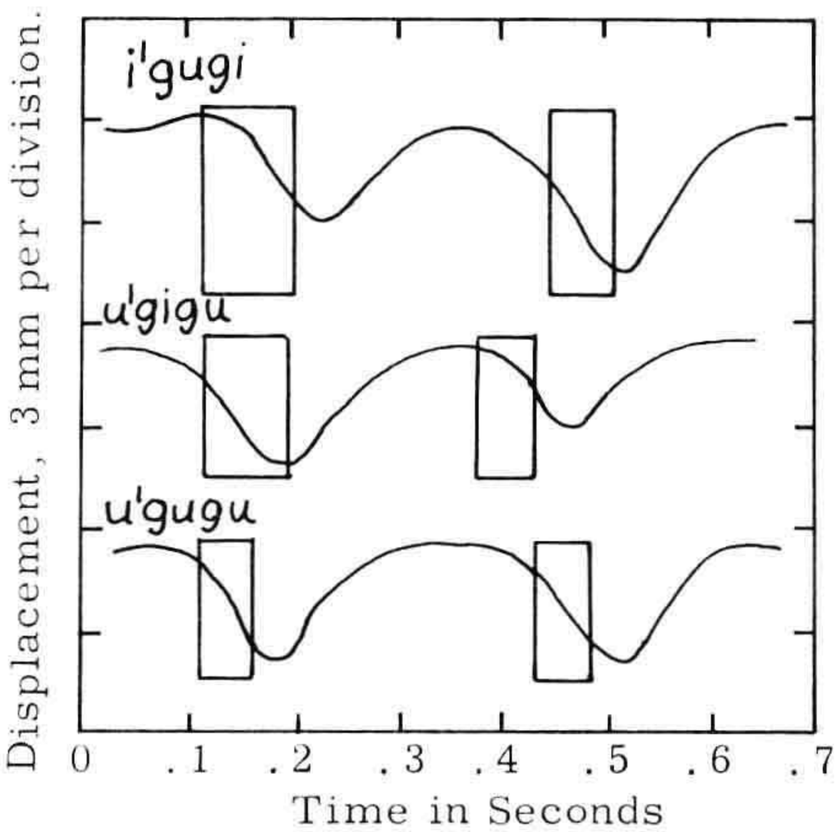


Figure 3. Forward directed perturbations during VgVgV utterances. Boxes indicate periods of palatal closure.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-1-5

Dynamic Analysis of the Vocal Tract

Joseph L. De Clerk
Research Scientist, U.S. Army Electronics Command,
Fort Monmouth, New Jersey
Jerry A. Selvaggi
Associate Professor, Gannon College,
Erie, Pennsylvania

Introduction

The vocal tract is modeled in sixteen quantized sections. Each section is represented in the time domain from cinefluorographic data. Webster's wave equation is forced to accept the dynamic area function. Radical changes in area functions are represented by delta functions.

Solution of Webster's Wave Equation Using Quantized Sections

We will assume that the dynamic behavior of the vocal tract can be described by the following form of Webster's equation under lossless conditions.

$$\frac{\partial^2 \alpha(x,t)}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 \alpha(x,t)}{\partial t^2} = -\frac{A'_x(x,t)}{A(x,t)}\frac{\partial \alpha(x,t)}{\partial x} \qquad (1)$$

where $\alpha(x,t)$ = position and time dependent potential ; c=speed of sound
$A(x,t)$ = cross-sectional area perpendicular to x
$A'_x(x,t) = \frac{\partial A(x,t)}{\partial x}$
$\alpha(\ell,t) = 0$ [at lips] ; $\frac{\partial \alpha(x,t)}{\partial x}\Big|_{x=0} = 0$ [at glottis]
ℓ = length of tract; initial conditions of time = 0

The area function, $A(x,t)$, is quantized in n sections, each section time dependent.

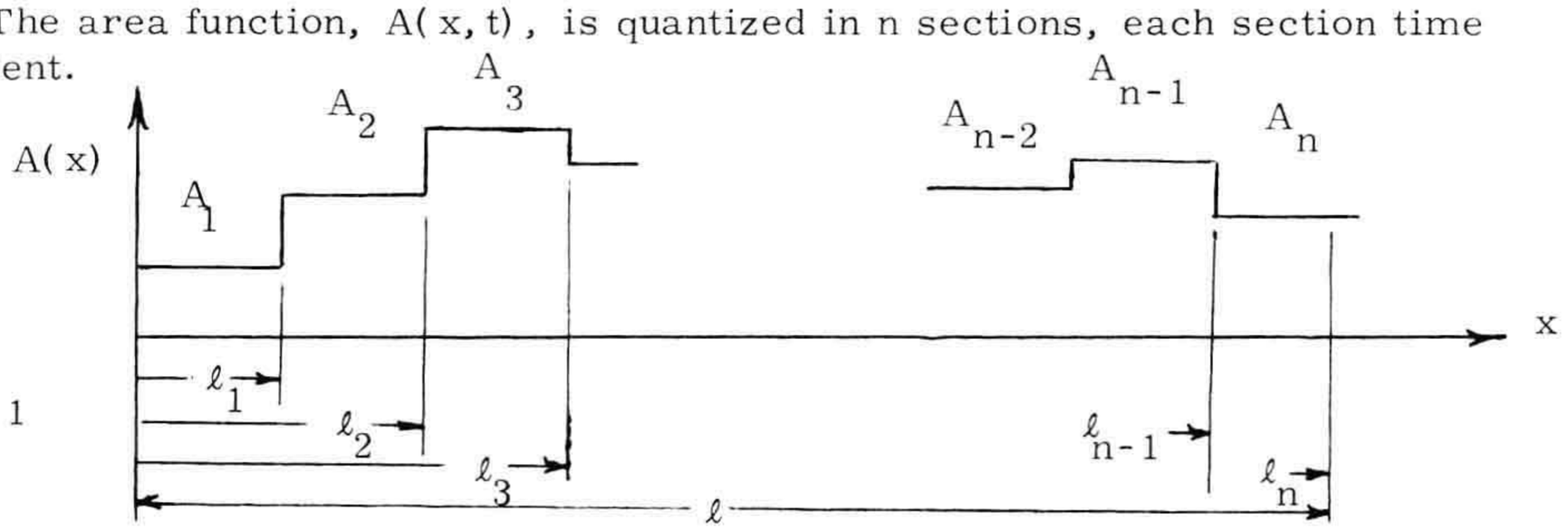


Fig. 1

Its mathematical description becomes

$$A(x,t) = A_1(t)\ \mu(\ell_1 - x) + A_2(t)\left[\mu(x-\ell_1) - \mu(x-\ell_2)\right] + A_3(t)\left[\mu(x-\ell_2) - \mu(x-\ell_3)\right]$$
$$+ \ldots A_n(t)\ \mu(x-\ell_{n-1}) \qquad (2)$$

where $\mu(x)$ = step function of magnitude one; zero for negative x, one for positive x

The definition of the delta function is $\int_{-\infty}^{\infty} \phi(x)\,\delta(x-\ell_1)\,dx = \phi(\ell_1)$

and $$\frac{d\mu(x-\ell_1)}{dx} = \delta(x-\ell_1)$$

Since $A(x,t)$ is a generalized function, its derivative becomes

$$\frac{\partial A(x,t)}{\partial x} = -\left[A_1(t) - A_2(t)\right]\delta(x-\ell_1) - \left[A_2(t) - A_3(t)\right]\delta(x-\ell_2)$$
$$- \left[A_{n-1}(t) - A_n(t)\right]\delta(x-\ell_{n-1}) \qquad (3)$$

The operation $\mu(x-\ell_n)\ \delta(x-\ell_n)\ \phi(x)$ is defined as zero, or

$$\mu(x-\ell_1)\ \delta(x-\ell_1)\ \phi(x) = 0 \qquad (4)$$

The operation $$\phi(x)\ \delta(x-\ell_1) = \phi(\ell_1)\ \delta(x-\ell_1) \qquad (5)$$

The term $\frac{A_x'(x,t)}{A(x,t)}$ is written as

$$\frac{A_x'(x,t)}{A(x,t)} = -\frac{\sum_{1}^{n-1}\left[A_n(t) - A_{n+1}(t)\right]\delta(x-\ell_n)}{A_{n+1}(t)} \qquad (6)$$

Iteration will be employed to solve Webster's wave equation.

Since the right hand member of EQ. (1) is a disturbing term, it will be forced to zero. Then the wave equation becomes

$$\frac{\partial^2 \varphi_0(x,t)}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 \varphi_0(x,t)}{\partial t^2} = 0 \qquad (7)$$

Employing the rules of transformation calculus (Laplace) twice, and invoking the initial and boundary conditions, the solution becomes

$$\varphi_0(s,x) = \phi(s,0)\ \cosh\ \frac{sx}{c} \qquad (8)$$

let $\phi(t,0) = \delta(t)$, then $\phi(s,0) = 1$ and

$$\varphi_0(s,x) = \cosh\ \frac{sx}{c} \qquad (9)$$

or the Green's function in the s, x domain. Then

$$\varphi_0(x,t) = \frac{1}{2\pi i}\int_{Br(s)} \cosh\ \frac{sx}{c}\ \epsilon^{st}\,ds \qquad (10)$$

The solution of (9) when x = ℓ results in the formants of the tract whose area function is constant. Also

$$\frac{\partial}{\partial x}\varphi_0(x,t) = \frac{1}{2\pi i}\int_{Br(s)} \frac{s}{c} \sinh\frac{sx}{c}\ \epsilon^{st} ds \qquad (11)$$

Substituting (6) , (11) in (1) the wave equation becomes

(12)

$$\frac{\partial^2\varphi_1}{\partial x^2}(x,t) - \frac{1}{c^2}\frac{\partial^2\varphi_1}{\partial t^2}(x,t) = \sum_1^{n-1} \frac{[A_n(t) - A_{n+1}(t)]\,\delta(x-\ell_n)}{A_{n+1}(t)} \cdot \frac{1}{2\pi i}\int_{Br(s)} \frac{s}{c} \sinh\frac{sx}{c}\ \epsilon^{st} ds$$

$$\text{Let } \sum_1^{n-1} \frac{[A_n(t) - A_{n+1}(t)]}{A_{n+1}(t)} = \sum_1^{n-1} f_n(t)\ \delta(x-\ell_n) \qquad (13)$$

Then (12) becomes

$$\frac{\partial^2\varphi_1}{\partial t^2}(x,t) - \frac{1}{c^2}\frac{\partial^2\varphi_1}{\partial t^2}(x,t) = \sum_1^{n-1} f_n(t)\ \delta(x-\ell_n) \cdot \frac{1}{2\pi i}\int_{Br(s)} \frac{s}{c} \sinh\frac{sx}{c}\ \epsilon^{st} ds \qquad (14)$$

Transforming on t and invoking complex convolution, (14) becomes

$$\frac{d^2\varphi_1}{dx^2} - \frac{s^2}{c^2}\varphi_1(s,x) = \sum_1^{n-1} \delta(x-\ell_n) \cdot \frac{1}{2\pi i}\int_{Br(\alpha)} F_n(\alpha)\ \frac{(s-\alpha)}{c} \sinh\frac{(s-\alpha)}{c}x\ d\alpha \qquad (15)$$

Invoking (5) , (15) becomes

$$\frac{d^2\varphi_1(s,x)}{dx^2} - \frac{s^2}{c^2}\varphi_1(s,x) = \sum_1^{n-1} \delta(x-\ell_n) \cdot \frac{1}{2\pi i}\int_{Br(\alpha)} F_n(\alpha)\frac{(s-\alpha)}{c} \sinh\frac{(s-\alpha)}{c}\ell_n\ d\alpha \qquad (16)$$

Transforming on x or $[x \rightarrow z]_{n-1}$ and imposing the boundary conditions, (16) becomes

$$z^2\varphi(s,z) - z\,\frac{-s^2}{c^2}\varphi(s,z) = \sum_1^{n-1} \left[\frac{1}{2\pi i}\int_{Br(\alpha)} F_n(\alpha)\frac{(s-\alpha)}{c} \sinh\frac{(s-\alpha)}{c}\ell_n\ d\alpha\right] \epsilon^{-\ell_n z}$$

The inverse transform becomes

(17)

$$\varphi_1(s,x) = \cosh\frac{sx}{c} + \sum_1^{n-1} \left[\frac{1}{2\pi i}\int_{Br(\alpha)} F_n(\alpha)\frac{(s-\alpha)}{c} \sinh\frac{(s-\alpha)}{c}\ell_n\ d\alpha\right] \cdot \frac{c}{s} \sinh\frac{s(x-\ell_n)}{c}\,\mu(x-\ell_n)$$

Let the n quantized sections be of equal length, or $\frac{\ell}{n}$. Employing (13) , each section is expressed in a Fourier cosine series. Let

$$f_1(t) = A_{10} + A_{11}\cos\omega_0 t + A_{12}\cos 2\omega_0 t + \ldots\ldots A_{1m}\cos m\omega_0 t$$

$$f_2(t) = A_{20} + A_{21}\cos\omega_0 t + A_{22}\cos 2\omega_0 t + \ldots\ldots A_{2m}\cos m\omega_0 t$$

$$f_{n-1}(t) = A_{(n-1)0} + A_{(n-1)1}\cos\omega_0 t + A_{(n-1)2}\cos 2\omega_0 t + A_{(n-1)m}\cos m\omega_0 t$$

Summing the result becomes $\sum_{n=1}^{n-1} f_n(t) = \sum_{n=1}^{n-1}\sum_{m=0}^{m} A_{nm}\cos m\omega_0 t$ (18)

The constant terms determine the static formant structure. The dynamic terms generate excursions about the static formant structure. Eq. (13) becomes

$$\sum_{n=1}^{n-1} f_n(t)\ \delta\left(x - \frac{n\ell}{n_0}\right) \sum_{n=1}^{n-1}\sum_{m=0}^{m} A_{nm}\cos m\omega_0 t\ \delta\left(x - \frac{n\ell}{n_0}\right) \quad (19)$$

where n_0 equals the number of quantized sections.

The solution of the wave equation after one iteration is Eq. (17) which becomes

$$\varphi_1(s,x) = \cosh\frac{sx}{c} + \sum_{n=1}^{n-1}\sum_{m=0}^{m} A_{nm}\left[\sinh\frac{ns\ell}{n_0 c}\cos\frac{nm\omega_0\ell}{n_0 c} - \frac{m\omega_0}{s}\cosh\frac{ns\ell}{n_0 c}\sin\frac{nm\omega_0\ell}{n_0 c}\right]$$

$$\cdot\left[\sinh\frac{s}{c}\left(x-\frac{n\ell}{n_0}\right)\ \mu\left(x-\frac{n\ell}{n_0}\right)\right] \quad (20)$$

The first term determines the formant structure for a constant area function; the series contributes both to static and dynamic excursions. Imposing the first boundary condition

$$\varphi_1(s,\ell) = \cosh\frac{s\ell}{c} + \sum_{n=1}^{n-1}\sum_{m=0}^{m} A_{nm}\left[\sinh\frac{ns\ell}{n_0 c}\cos\frac{nm\omega_0\ell}{n_0 c} - \frac{m\omega_0}{s}\cosh\frac{ns\ell}{n_0 c}\sin\frac{nm\omega_0\ell}{n_0 c}\right]$$

$$\cdot\left[\sinh\frac{s}{c}\left(\ell-\frac{n\ell}{n_0}\right)\ \mu\left(\ell-\frac{n\ell}{n_0}\right)\right] = 0 \quad (21)$$

The solution results in eigenvalues or formants. The static case becomes

$$\varphi_1(s,\ell) = \cosh\frac{s\ell}{c} + \sum_{n=1}^{n-1} A_{n_0}\sinh\frac{ns\ell}{n_0 c}\sinh\frac{s}{c}\left(\ell-\frac{n\ell}{n_0}\right) = 0 \quad (22)$$

Accuracy increases with successive iterations. The complete solution results after (n-1) iterations.

Area function data of the syllable /bæd/ was quantized into 16 equal sections. The total length of the tract was 15.1 cms. Each section was plotted in the time domain. Invoking Eq. (13), the behavior of the tract from glottis to lips was represented by Fourier series. Only the first two terms of the Fourier series were retained. Invoking Eq. (21) the solution becomes

$$\varphi(z) = 200\cosh^8 z + 10.45\cosh^7 z - 405\cosh^6 z - 24.2\cosh^5 z + 255\cosh^4 z$$

$$+ 15.9\cosh^3 z - 51\cosh^2 z - 2.64\cosh z + 1.497 = 0 \quad (23)$$

where $z = \frac{2(15.1)\ s}{16\ c}$

The transfer function is the reciprocal of Eq. (23), or $T(z) = \frac{1}{\varphi(z)}$ (24)

The poles of Eq. (24) become the formants of the tract. A number of calculated values in hertz are: 340, 1660, 2900, 4140, 5100, 7540, 11000, 12200, 13300 and 14400.

Acknowledgement

The authors would like to acknowledge the computational assistance of T. McBride.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-1-7

A Study of Dynamic Palatography

Osamu Fujimura, Sadao Shibata,

Shigeru Kiritani and Zyun'ici Simada

Research Institute of Logopedics and Phoniatrics,
Faculty of Medicine, University of Tokyo

Cotoji Satta

The National Center of Hearing and Speech Disorders

A new experimental method has been devised for the observation of dynamic changes in the contact between the tongue and the roof of the mouth during articulatory processes.

Procedures

Sixty-four pairs of small silver electrodes (1-mm in diameter) are implanted into a thin plastic artificial palate, the thickness of which is approximately 1-mm. One of each pair is connected to a Wien-bridge oscillator for its assigned frequency, the other to a busline for a common output.

Sixty-four oscillators covering a frequency range from 5000- to 15650-Hz with a space of 150-Hz are used to generate palatal signals. The palatal signals are recorded on one track of magnetic tape, and the speech signal is recorded on the other track. A block diagram is shown in Fig. a.

The signals played back in the half speed are mixed and fed into a sound-spectrograph. For the lower frequency range of the spectrogram, the speech signal is displayed by use of the wide-band filter. The contact signals from different points are displayed as horizontal lines in the upper part (between

2500-Hz-7825-Hz) of the same spectrographic display.

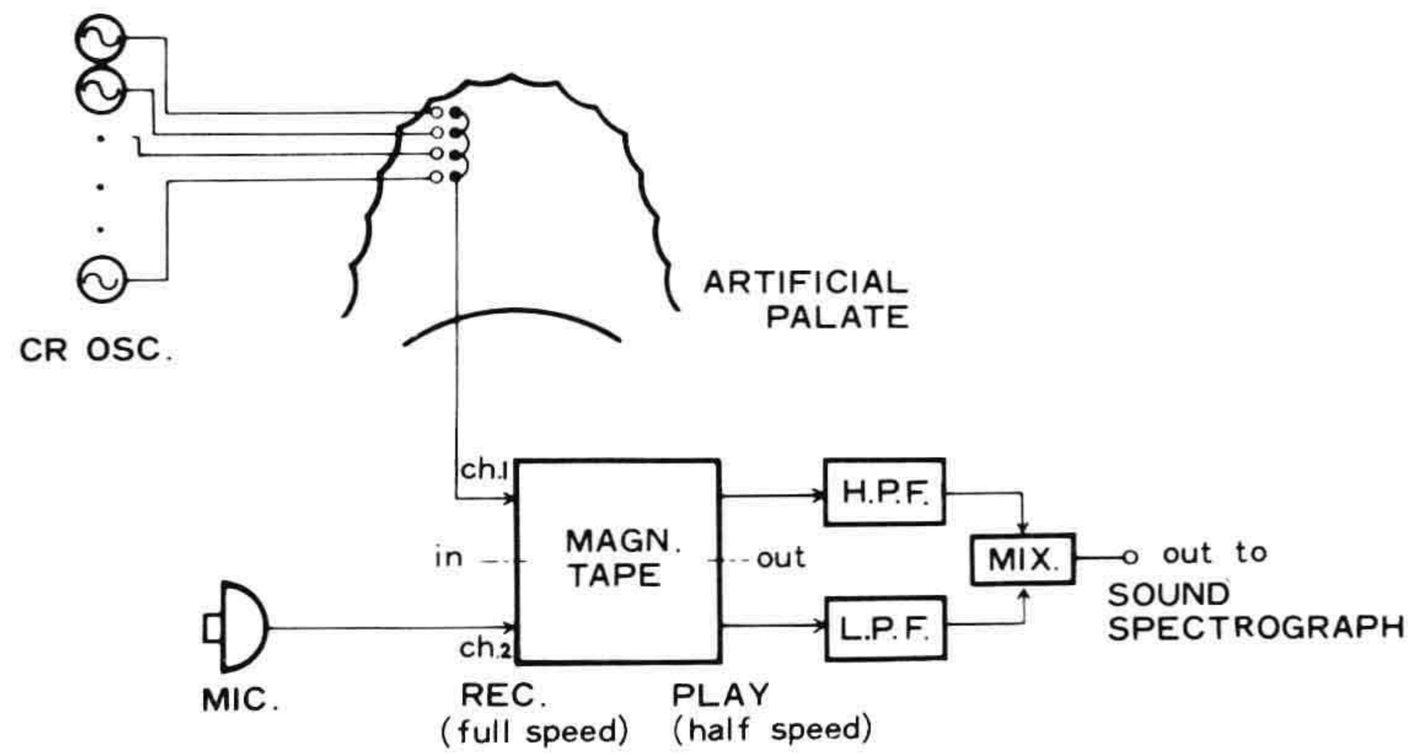


Fig. a

Special care is taken for the best arrangement of the palatal signals on the recording paper, viz. the connection of the oscillators to the palatal electrodes, in accordance with a specific point of interest of the study. In one case, for example, sixty-four points are divided into four groups, in each of which a progressively higher frequency is assigned to every point in the order of increasing distance from the last molar teeth of the right side to that of the left (Fig. b).

In the preliminary experiment reported in the followings, the subject is instructed to utter meaning-less bi-syllabic (VCV) Japanese words.

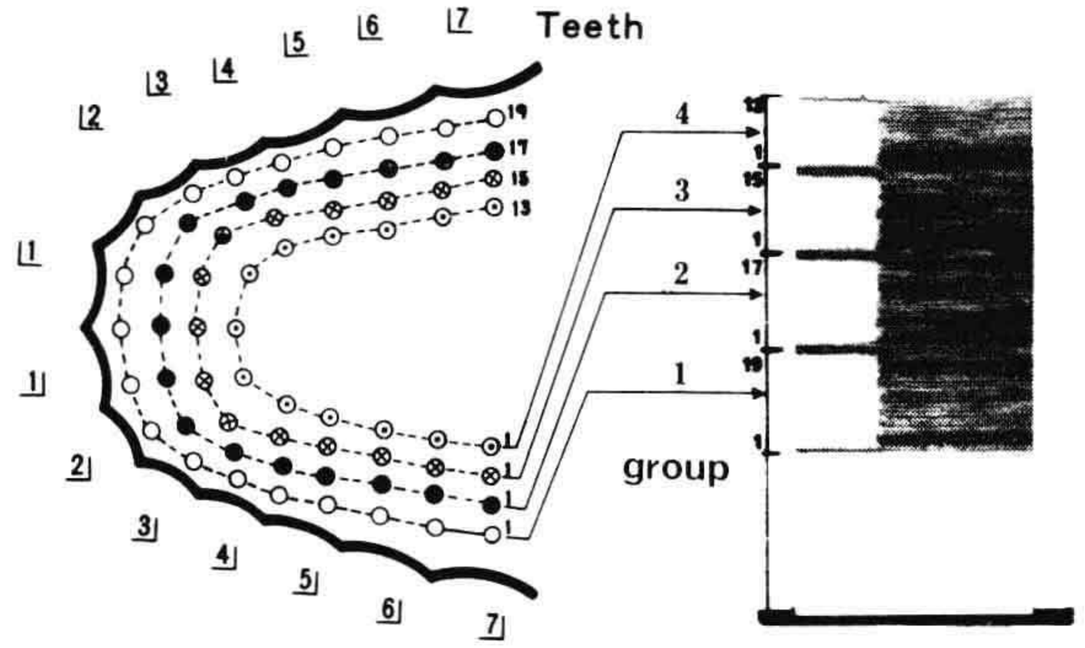


Fig. b

Results

The patterns illustrated in Fig. c show details of the manner of contact in selected Japanese consonants.

[t]: Some contacts are clearly seen in all groups except for the fourth group. In the initial phase of closure, the contact in each group (cf. Fig. b) begins at the distal (molar teeth) region and successively reaches the medial (incisor teeth) region. The lower points on the palate show earlier contacts than the higher ones (Compare group 3 to group 1 in Fig. c).

In contrast, the release proceeds from the medial to the distal and from the higher to the lower. The acoustic release approximately coincides with the release of the contact in the central portion of the palate. A close inspection of the record of the contacts reveals details of the articulatory movements.

[d]: The contact pattern appears to be similar to that of [t], except that the area of contact is significantly smaller. In this sense the articulation of [d] is lax compared to that of [t].

[r]: The pattern is entirely different from that of all other consonants and it reveals characteristics of a flapping articulation.

[s]: There is no contact in the central electrodes of any group. This shows that a narrow central grooving of the tongue takes place.

[ʃ]: Compared with the pattern of [s], a wider space in the central portion remains free from contact, while the contacts are seen in the higher palatal positions.

[ts], [tʃ]: In the initial phase of the articulatory movement, the manner of contact is similar to that for [t]. The temporal pattern of release, on the other hand, seems similar to that of [s] and [ʃ], respectively.

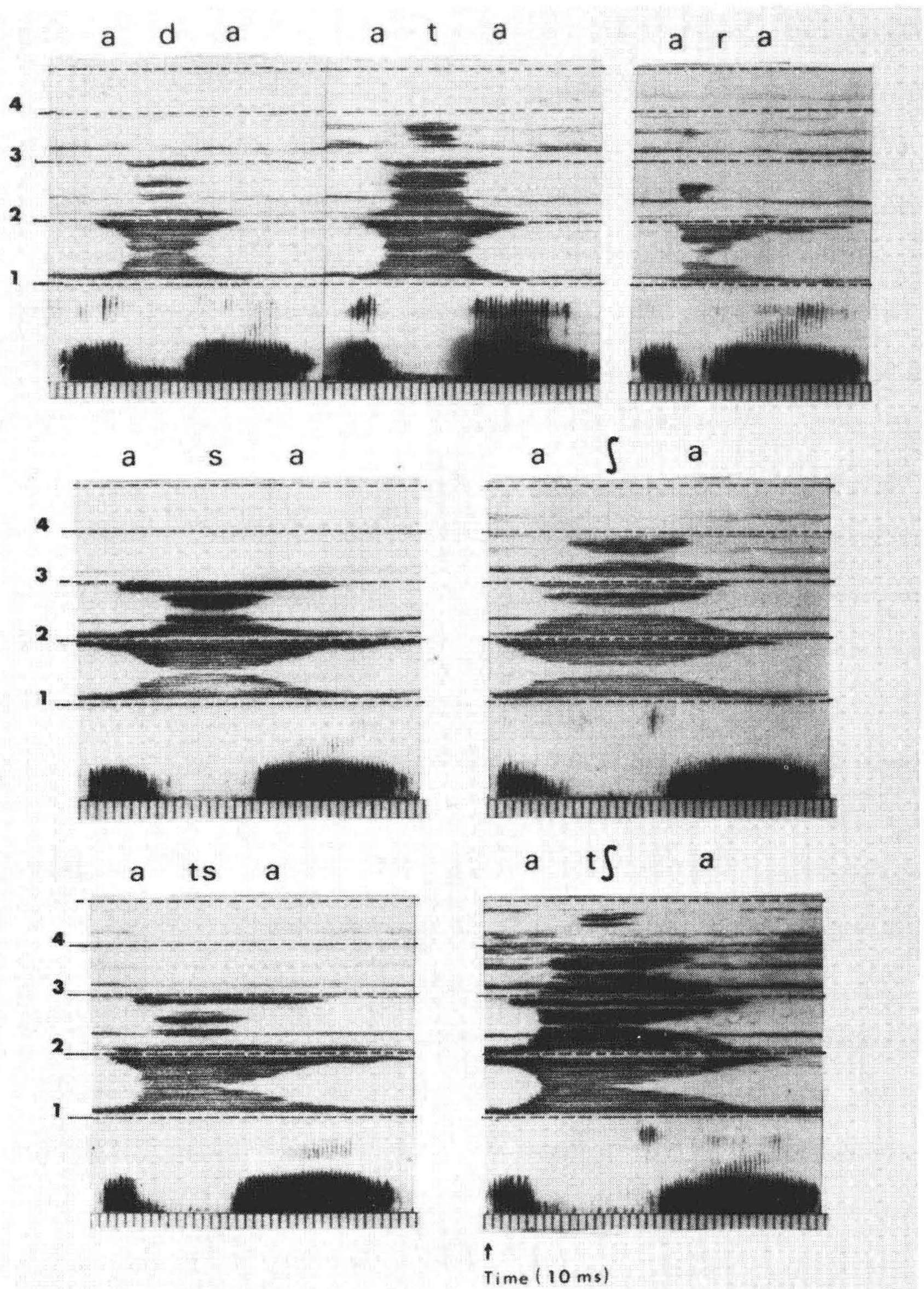


Fig. c Records of the dynamic palatography

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-1-8

An Electromyographic Study of Articulatory Movements

Shigeru Kiritani, Hajime Hirose and Masayuki Sawashima

Research Institute of Logopedics and Phoniatrics,

Faculty of Medicine, University of Tokyo

An electromyographic study is made on the activities of some articulatory muscles with special reference to the speed of utterance. One of the most interesting problem in the study of articulatory movements is how the sequence of motor commands which are inherent to each phoneme is organized to bring about an articulatory gesture as a whole. The aim of the present study is to investigate the time relationship between the muscle activity and corresponding acoustic event in voluntary speech.

As the object muscles, the orbicularis oris and the anterior belly of the digastric are selected in the present study.

The orbicularis oris is known to be responsible for articulatory gestures of rounding, protrusion and closure of the lips. Among five Japanese vowels /a/ /i/ /u/ /e/ /o/, /o/ is a rounded vowel and is associated with an activity of the orbicularis oris.

The digastric is considered to be responsible for opening the jaw. As for the Japanese vowels, the digastric shows marked activities for /a/ and /o/, in which jaw opening is wider than in other three vowels.

Experimental Procedures

In the present study, non-sense Japanese words are used as test words. Each word is selected so as to contain the pertinent Japanese vowel for the activity of either the orbicularis or the digastric. Thus, /kaetori/ is used in the series of experiment on the orbicularis oris, while /uriarito/ is used in that on the digastric. The muscle activity is expected for the underlined vowel in each test word.

The subject sets three different speeds of utterance designated as slow, moderate and fast, respectively. For each speed, the word is uttered in isolation and is repeated ten times in series keeping the speed of utterance as constant as possible. In the series of fast speed, the subject is instructed to articulate in his maximal speed of speech.

The subject is placed in an electrically shielded room. A special monopolar needle electrode is used throughout the present study. The needle electrode is made of a Japanese acupuncture needle, the diameter of which is about 0.2 mm with a very sharp tip. It is insulated except for the tip measuring approximately 0.5 mm. The needle of this type is so thin and pliable that the insertion of the needle does not give much discomfort to the subject and does not affect normal articulatory movements.

The E.M.G. signals and the acoustic signals are recorded simultaneously on a two channel tape recorder. Fig. 1 shows examples of raw recordings of the E.M.G. and acoustic signals for each test word.

Fig. 2 shows a block diagram of the system employed to obtain averaged E.M.G. Data. The tape is played back continuously for one series of utterance. The E.M.G. signal is fed into a small digital data processing computer. A trigger pulse is generated at the onset of the acoustic signal of each utterance. The E.M.G. signal is processed for one second following

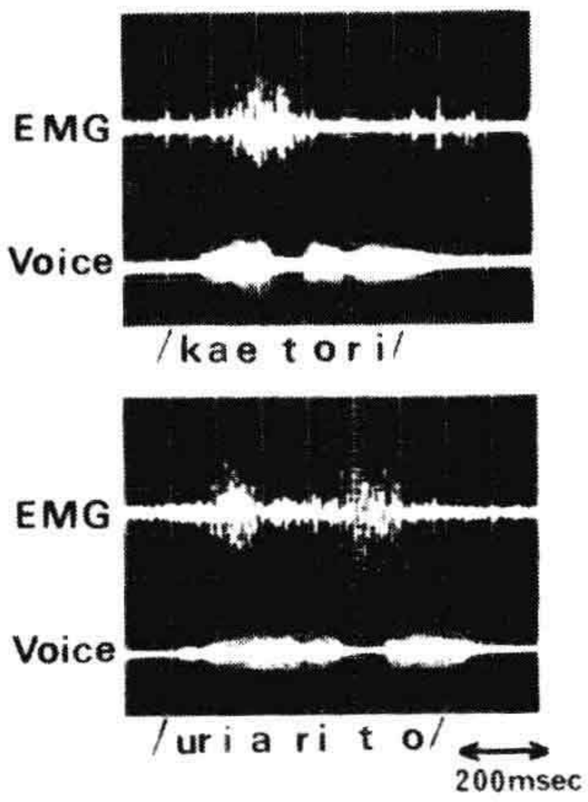


Fig. 1 Examples of raw recordings

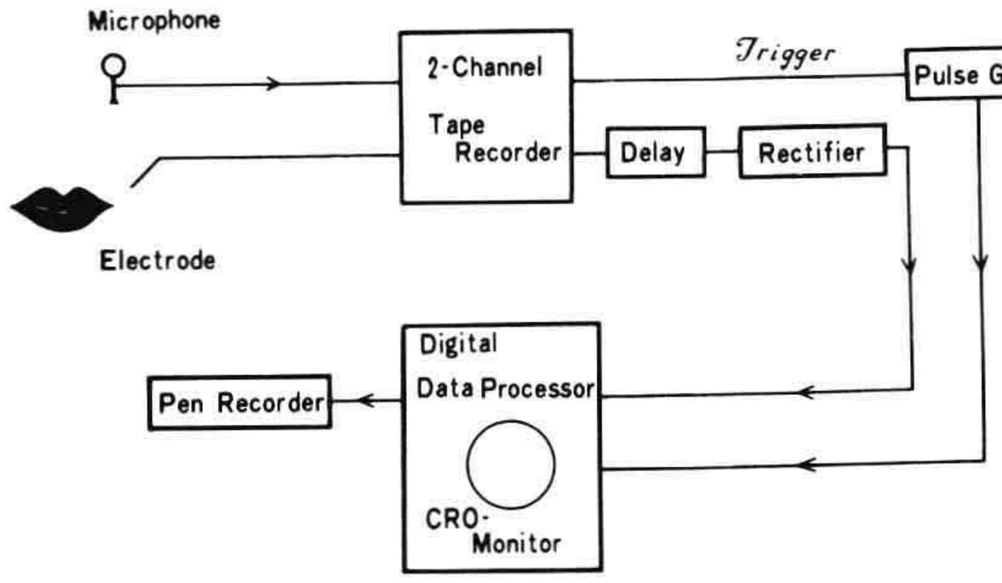


Fig. 2 Block diagram of the system employed to obtain averaged E.M.G. data

the trigger pulse and is integrated for every 10 msec to obtain 100 sampled values in one second. The values are averaged for eight utterances out of ten in each series. Averaged E.M.G. data are ei ther displayed on a monitor oscilloscope or recorded by a pen recorder.

The E.M.G. data averaged for eight utterances still show some irregularity. Nevertheless its peak position can be determined within a range of 20 msec in most cases. The time interval between the onset of the acoustic signal and the representative acoustic event is measured on the soundspectrogram for each utterance. The values are averaged and the reference points of the acoustic signals to the E.M.G. are decided.

Results and Remarks

Fig. 3 shows averaged E.M.G. data in three different speeds of utterance with reference to the acoustic signals. It is noted that in a slow utterance the peak of the averaged E.M.G. signal is found approximately at the beginning

of the acoustic signal of the corresponding phoneme, while in a fast utterance it is at the preceding phoneme, i.e. /e/ in /kaetori/ and /i/ in /uriarito/.

A possible interpretation of these results is that there is a comparatively constant time interval between the peak of the E.M.G. activity and the center of the corresponding acoustic event irrespective of the speed of utterance. As shown in the examples in Fig. 3 the time interval is found to be almost the same for both muscles examined and is approximately estimated at 110 msec.

For further analysis of the characteristics of the time interval, direct observations of movements of the articulatory organs associated with the E.M.G. recording are in schedule.

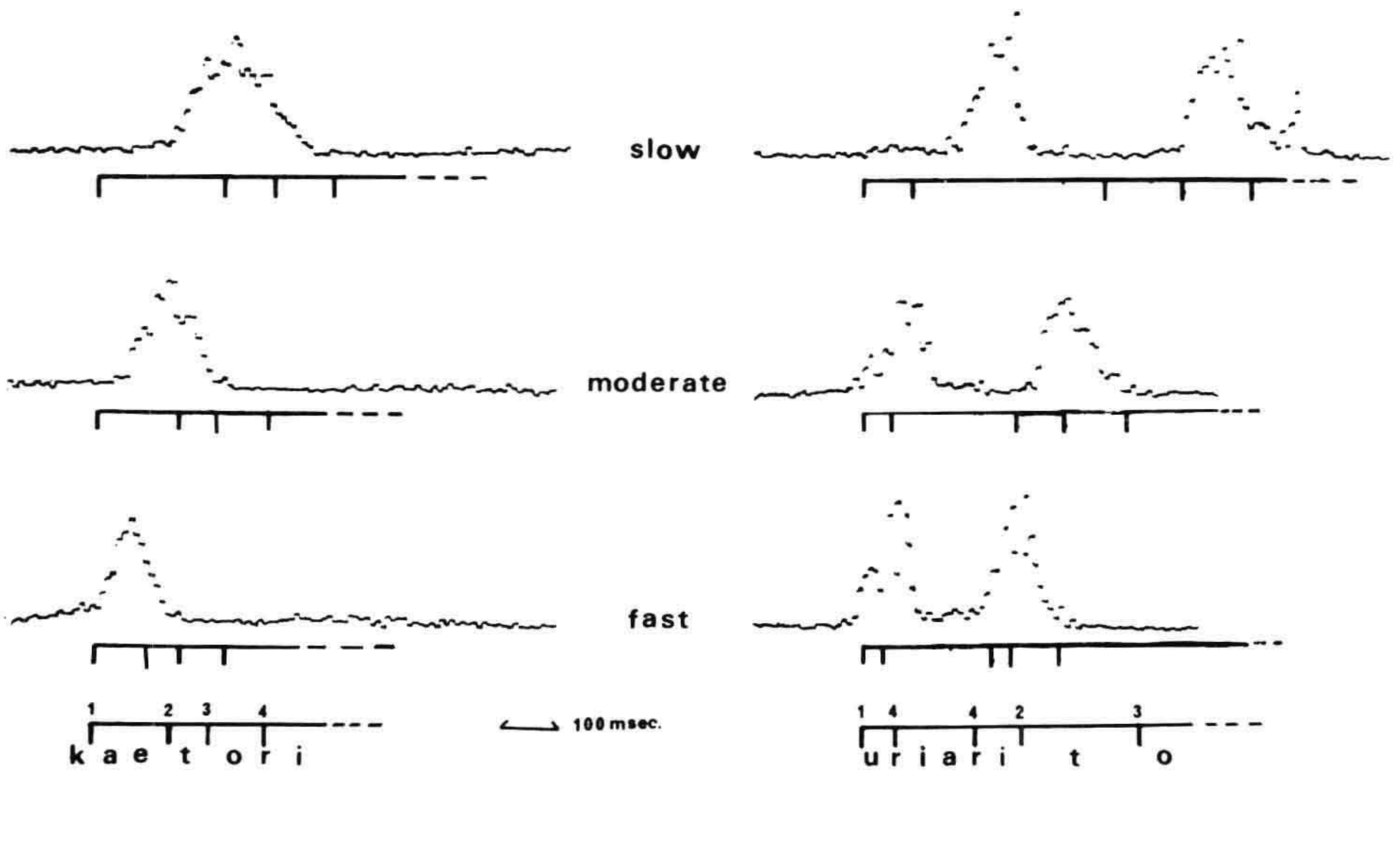


Fig. 3 Averaged E.M.G. data with reference to the acoustic signals.
1. voice onset 2. implosion of [t]
3. explosion of [t] 4. release of flapped [r]

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-1-9

Temporal Organization of Syllable Production

Björn E.F. Lindblom

Ph.D., Department of Speech Communication,

Royal Institute of Technology (KTH),

Stockholm, Sweden

The present paper draws attention to the classical phonetic observation demonstrated for many languages that under identical conditions the acoustic duration of an open vowel tends to be longer than that of a close vowel. An explanation of this fact is proposed in terms of a simplified model of articulatory coordination. The results obtained indicate that an articulatory model of speech production that is to predict the durational variance of vowels in a natural way, should incorporate the jaw as a separate component.

The open-close dimension of vowels is examined in terms of jaw opening. Good agreement with classifications based on tongue height such as the traditional vowel quadrilateral is obtained when the position of the mandible in long and short Swedish vowels is plotted against the classes of front-rounded, front-unrounded, and back-rounded vowels. When vowel duration in the context of bilabial stops is plotted against mandibular position a positive correlation can be observed. The lower the jaw the longer the vowel.

A dynamic model of lip and mandible coordination is constructed that gives an articulatory interpretation of the dependence of vowel duration on jaw position. The model in which the lips and the mandible are represented by damped spring-mass systems permits the derivation of the course of the midsagittal separation of the lips, $s(t)$ for [bVb] syllables. In terms of this parameter vowel duration

is defined as the interval during which $s > 0$ and can be determined from

$$s(t) = x_u(t) - x_\ell(t) - x_j(t) - x_o$$

This equation is by definition valid only when $x_u(t) - x_\ell(t) - x_j(t) - x_o \geq 0$. For $x_u(t) - x_\ell(t) - x_j(t) < 0$, $s(t) = 0$. The term x_u is the labiomuscular displacement of a point on the upper lip, x_ℓ is the labiomuscular displacement of a point on the lower lip and x_j stands for the jaw-dependent depression of the lower lip. The term x_o is the criterion of closure onset or release. This model is used to compute the hypothetical course of $s(t)$ in [bVb] syllables to which we assign various degrees of vowel or jaw opening.

The model predicts that vowel duration should increase as a function of the mandibular excursion for the vowel because the mandibular off-glide movement from the vowel into the final consonant progresses at such a rate that the contact between articulators for the consonant is delayed in the context of the open vowel. A further implication is that the durational variance of open and close vowels is not necessarily a characteristic of the input control signals to the model. It is compatible with invariant vowel duration at the level of control and can be generated simply as a result of the superposition of jaw movement on the opening and closing gestures of the lips.

A procedure was developed to record lip and jaw movements continuously and in synchrony with the speech signal. A number of utterance pairs differing minimally with respect to the degree of vowel opening were recorded using this method. The results of an analysis of this material make it appear likely that it is the dynamic behavior of the mandible that gives rise to the dependence of acoustic vowel duration on the degree of vowel opening.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-1-10

Sur la Production des Voyelles par des Locuteurs Hommes et Femmes

CARRE René, Chargé de Recherche au C.N.R.S.

LANCIA Roland, Professeur

Laboratoire de la Perception et de la Communication
Ecole Nationale Supérieure d'Electronique - Grenoble

WAJSKOP Max

Directeur de l'Institut de Phonétique
Université Libre de Bruxelles

Introduction

Il est admis que les écarts notables existant entre les fréquences des formants de voyelles émises soit par des hommes, soit par des femmes (1) proviennent de différences anatomiques, dans la longueur du conduit vocal notamment (2). Ce dernier est, en effet, plus court (de l'ordre de 2cm) chez la femme, et cette réduction affecte plus la longueur du pharynx que celle de la cavité buccale. D'autres différences portant, par exemple, sur les sections des constrictions doivent être prises en considération.

Par ailleurs, la fréquence de mélodie est, en moyenne, pour une femme, supérieure d'une octave à celle d'un homme.

Cependant, pour un homme par exemple, la mélodie peut atteindre des valeurs admises pour la voix de femme. Dans ce cas, les fréquences des formants de la voyelle émise évoluent-elles aussi vers celles des formants de la voyelle féminine correspondante ? Si oui, nous devons poser comme hypothèse qu'intervient alors une modification adéquate de la forme du conduit vocal.

Nous savons que notre larynx est mobile et que son déplacement est fonction de la fréquence vibratoire des cordes vocales (3). Existe-t-il une relation entre ce phénomène et l'hypothèse précédente ?

Pour répondre à cette question, nous avons effectué des expériences de simulation puis mesuré les fréquences des formants de voyelles émises par un locuteur modifiant sa fréquence fondamentale.

Etude du conduit vocal par simulation

A l'aide d'un simulateur du conduit vocal (4) nous avons simulé sept formes différentes du conduit vocal correspondant chacune à la production d'une voyelle déterminée. Pour définir ces formes, nous nous sommes inspirés des travaux référenciés 4, 5 et 6.

Après avoir noté, dans les différents cas, les fréquences F_1 et F_2 des deux premiers maxima (donc des deux premiers formants) des courbes de réponses du simulateur, nous avons court-circuité deux cellules de ce dernier, représentant 2cm du conduit, au niveau du début de la cavité du pharynx côté larynx. Cette réduction correspond approximativement au déplacement du larynx d'un homme lorsque la fréquence fondamentale augmente d'une octave. Nous avons noté les nouvelles valeurs F_1' et F_2' des fréquences des formants et calculé les écarts relatifs $k_1=(F_1'/F_1)-1$ et $k_2=(F_2'/F_2)-1$.

Les coefficients k_1 et k_2, ainsi obtenus pour les voyelles (ɛ), (e), (i), s'accordent avec les écarts relatifs, mesurés par FANT (2) ou PETERSON (1), entre les fréquences de formants des voyelles naturelles homologues féminines et masculines.

Par contre, pour les coefficients k_2 des voyelles du type postérieur (u), (o), (ɔ), (ɑ), l'accord est médiocre. Mais pour ces voyelles, une élévation du larynx lequel possède une certaine rigidité, doit provoquer une modification de la constriction principale, proche, dans ce cas, du larynx. Pour les voyelles (o), (ɔ), (ɑ), la constriction va s'élever légèrement; pour la voyelle (u), la constriction va se resserrer (voir les coupes du conduit vocal proposées par FANT réf.7). Sur la base de ces remarques, nous avons, sur notre simulateur soit déplacé la constriction de 1cm, soit réduit la section de cette dernière. Toutes ces modifications conduisent finalement, pour l'es-

sentiel, à raccourcir la longueur du conduit vocal, la réduction etant plus sensible sur la longueur du pharynx que sur celle de la bouche.

Le conduit vocal de l'homme tendrait donc à se transformer en un conduit vocal de femme lorsque la mélodie croît. Les résultats obtenus pour k_1 et k_2 après nos dernières modifications sont en bon accord avec les mesures de FANT et PETERSON (figures 1 et 2).

Fig. 1

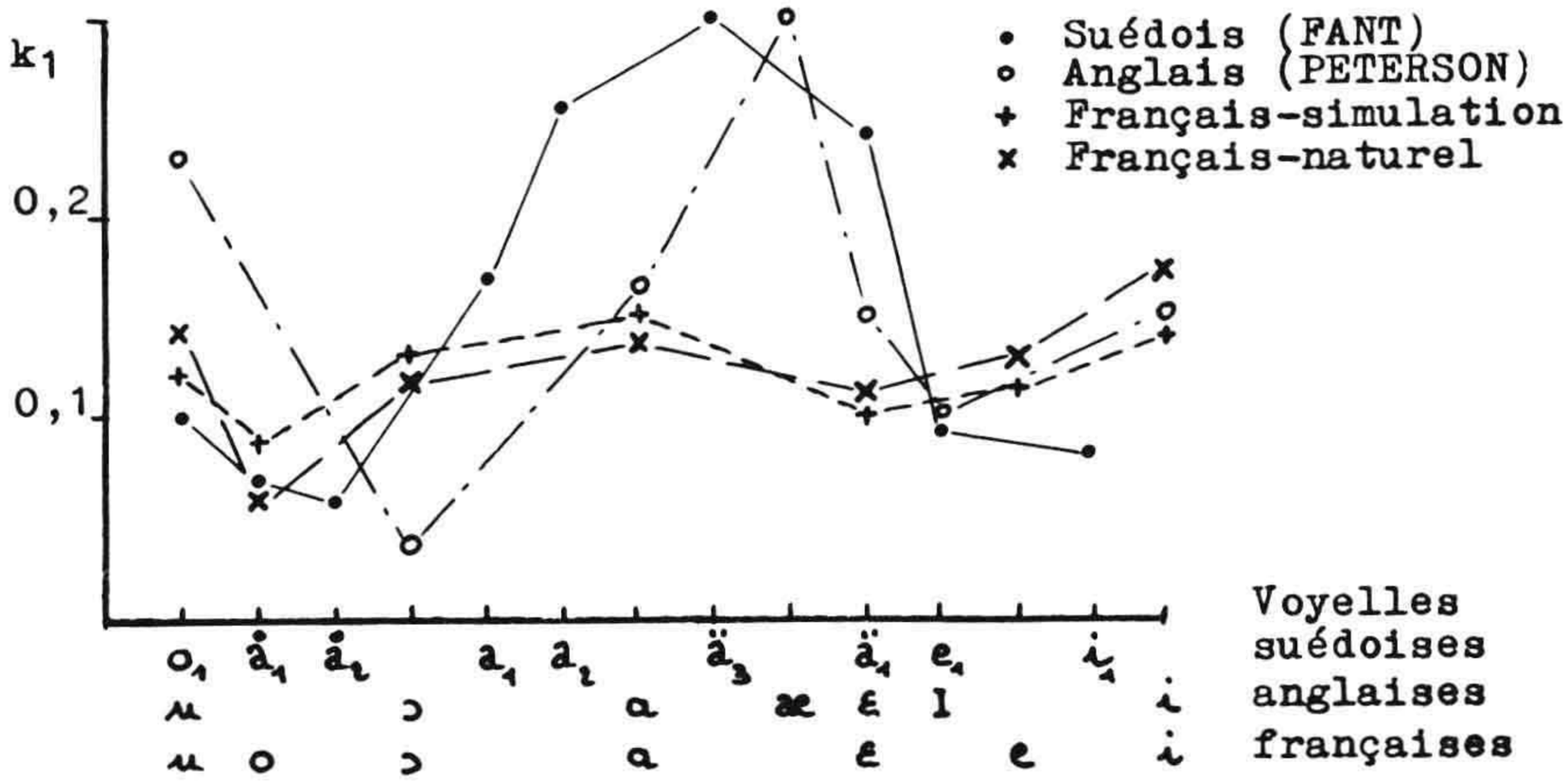
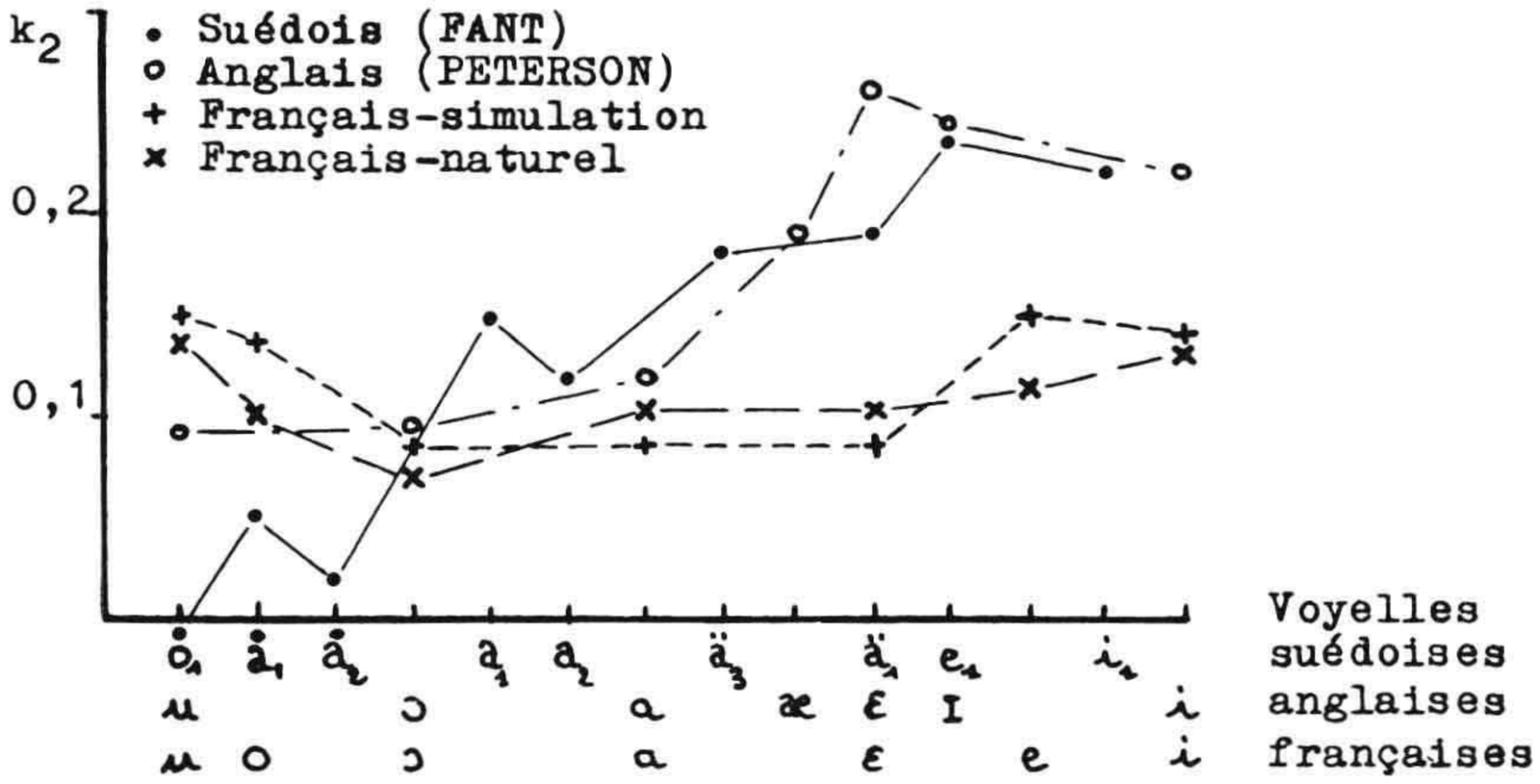


Fig. 2



Analyse spectrale de voyelles naturelles

Nous avons analysé les sept voyelles précédentes prononcées à

mélodie variable (la mélodie variant de 100 Hz à 200 Hz), par un locuteur masculin. Les fréquences des formants, F_1 et F_2 pour une mélodie de 100 Hz, F'_1 et F'_2 pour une mélodie de 200 Hz, ont été déterminées à l'aide de courbes standards. Nous avons calculé et représenté figures 1 et 2, les valeurs des écarts relatifs : $k_1=(F'_1/F_1)-1$; $k_2=(F'_2/F_2)-1$. L'accord entre ces écarts et ceux obtenus par simulation est satisfaisant.

Conclusion

Nos premières expériences montrent que le conduit vocal de l'homme tend à se transformer en un conduit vocal de femme lorsque la mélodie augmente. Réciproquement, le conduit vocal de la femme tend-il à se transformer en conduit vocal d'homme lorsque la mélodie diminue ?

Pour préciser ces différents points, nous nous proposons d'effectuer de nouvelles expériences en diversifiant les locuteurs (hommes et femmes) et en étudiant le conduit vocal à l'aide des rayons X.

Références

1 - G.E. PETERSON, H.L. BARNEY - Control methods used in a study of the vowels - J. Acoust. Soc. Am., 24, 175-184 (1952)

2 - G. FANT - A note on vocal tract size factors and non-uniform F-Pattern scalings - S.T.L., Q.P.S.R., 4/1966, Stockholm

3 - G.E. ARNOLD - Morphology and physiology of the speech organs - Manual of Phonetics, pp. 31-64 (éd. L. Kaiser)

4 - K.N. STEVENS, S. KASOWSKI, G. FANT - An electrical analog of the vocal tract - J. Acoust. Soc. Am., 25, 734-742 (1953)

5 - T. CHIBA, M. KAJIYAMA - The vowel : its nature and structure Tokyo, Kaiseikan (1941)

6 - G. FANT - Acoustic theory of speech production - S'Gravenhage, Mouton (1960)

7 - G. FANT - Formants and cavities - Proc. of the 5th Cong. of Phonetics Sciences, Münster (1964).

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

P-0-1

Pitch Extraction and Display for the Deaf

Ladislav Dolansky
Professor of Electrical Engineering,
Northeastern University, Boston, Mass.

Introduction

Because of the inadequate or altogether missing auditory feedback, the speech of deaf persons is often very unnatural, not only with respect to articulation, but also with respect to other features which are hard to monitor, such as intonation.

I. The Problem

This paper is concerned with the problem of the development of electronic equipment for the extraction of the pitch signal and its display in the form of a visual pattern, and with evaluation of the improvement obtained with deaf adults.

II. Approach

Nine deaf adults were used as experimental subjects. It was attempted to teach them various artificial or speech-derived intonation patterns, using a visual pitch-signal display. The following signals were used in consecutive sessions: (1) constant-pitch signal at one of the frequencies listed in Table I, (2) signals with a step change between these frequencies, either up or down (see Fig. 2a), (3) signal derived from three-word phrases of conversational speech, (4) extreme rising or falling inflections, such as are encountered in "oh", "well", etc. Towards the end of the training, the subjects were given only delayed visual feedback. The subjects' responses were recorded for later study.

III. Experimental Equipment

The pitch-signal extractor, indicated by the dashed enclosure in Fig. 1 is based on the envelope-detection idea used in the instantaneous pitch-period indicator of Dolansky[2], which was later modified by Anderson[1], and more recently by Phillips[3]. The rest of the equipment shown in Fig. 1 is used as follows. A manually started Ampex recorder gives a pre-recorded signal. Each sample is preceded by a short 10 kHz signal, which starts a sequence of time delays $\Delta\tau_1$, $\Delta\tau_2$, and $\Delta\tau_3$. During

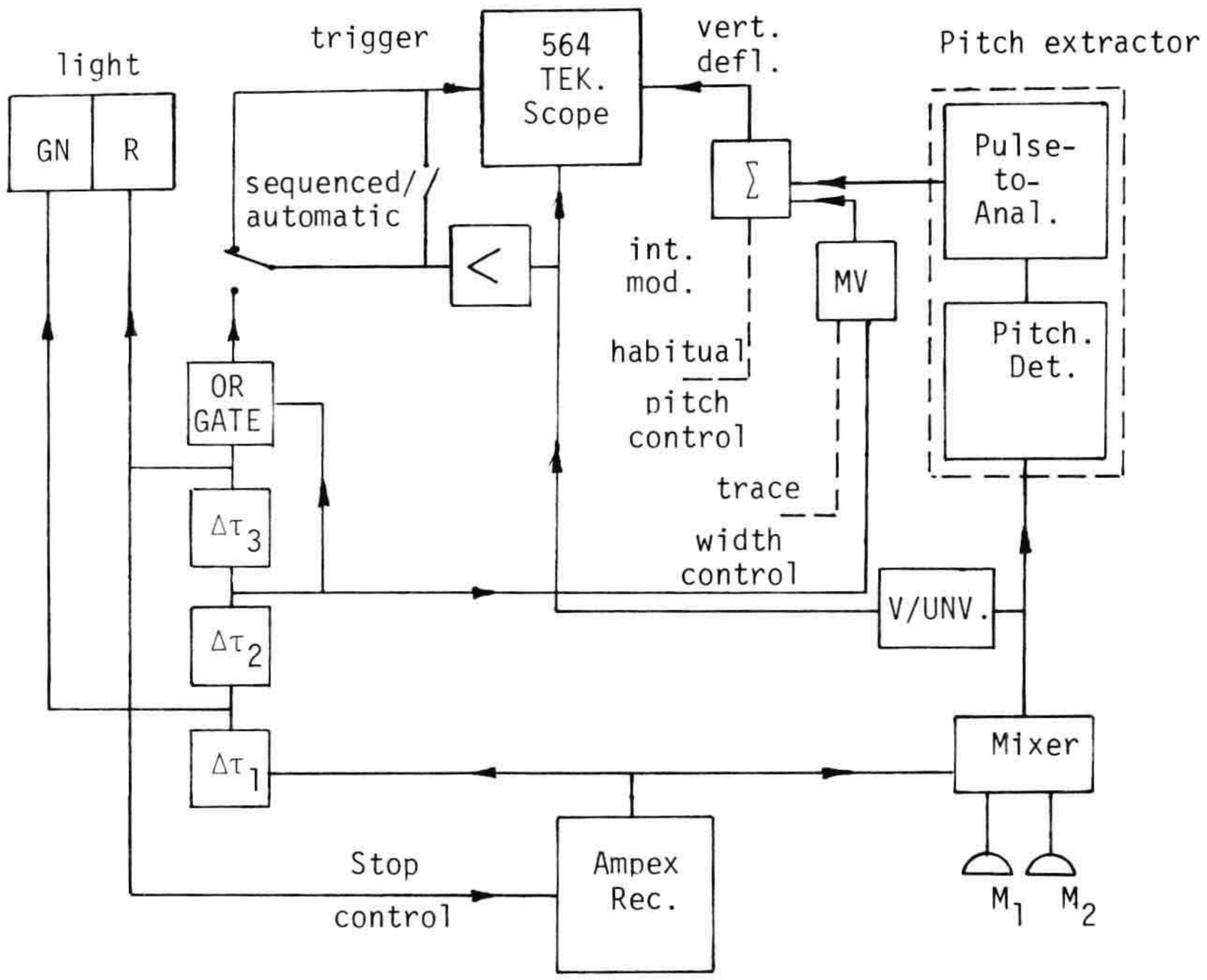


Fig. 1. Experimental pitch-extraction and display equipment.

TABLE I. Frequencies(Hz)used for constant or step patterns.

Pitch	Male	Female
Low	130	220
Medium	155	261
High	185	311

$\Delta\tau_1$, the green light (GN) indicates that a stimulus signal will soon appear on the screen of the oscilloscope. During $\Delta\tau_2$, a multivibrator is activated; the latter superposes a rapid square wave of a given height onto the intonation contour (see Fig. 2(b)), giving the subject the limits between which his own utterances should be held. During $\Delta\tau_3$, a red light (R) indicates that the subject's utterance should begin. An OR gate starts the oscilloscope sweep at the right instant for the trace to appear on

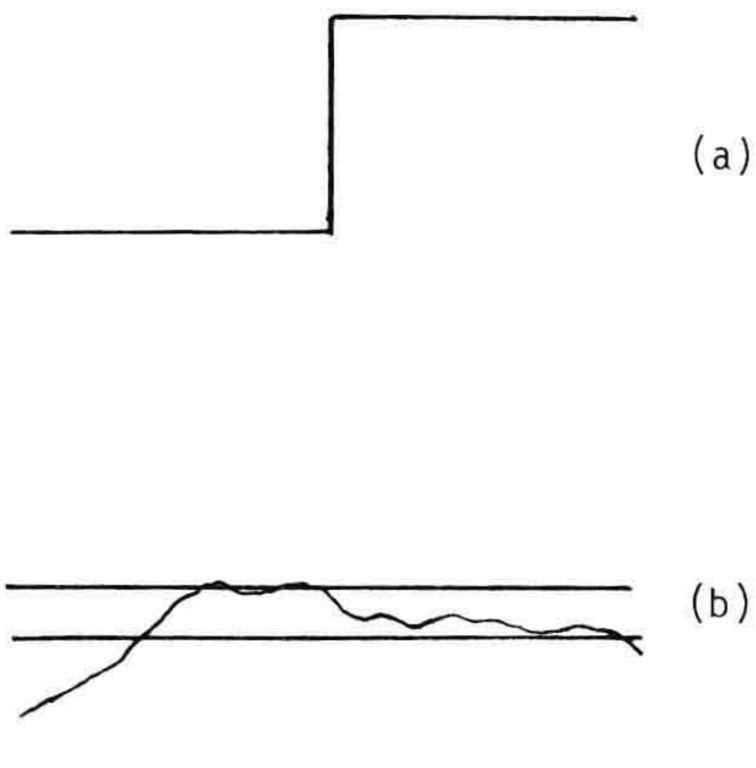


Fig. 2. (a) stimulus pitch-contour pattern with a step; (b) constant frequency stimulus pattern with indicated bounds, and subject's response.

TABLE II. Percent of responses that listeners judged post-test better than pre-test, taken for each subject.

Subject	Group 1	Group 2	Group 3
1	33.65	29.64	26.15
2	62.05	46.15	53.54
3	78.67	78.82	79.80
4	59.49	37.41	45.54
5	58.79	38.69	45.15
6	58.79	51.81	52.62
7	51.65	37.10	44.46
8	45.05	64.71	59.69
9	51.10	54.75	48.77

the screen. For unvoiced utterances, the voiced-unvoiced detector (V/UNV) insures that the trace is blanked out during the unvoiced interval. In addition, at the beginning of $\Delta\tau_3$, the recorder is stopped. The oscilloscope is capable of storing both stimulus and response patterns indefinitely, i.e. until such time as both patterns are erased by manual control.

IV. Experimental results

Examples of visual displays are given in Fig. 2 a,b. The display shown in Fig. 2 corresponds to a sinusoidal signal whose frequency suddenly changes from the low to the high frequency (see Table I). Fig. 2 shows a constant-frequency stimulus pattern onto which a rapid square wave was superposed so that a double line results. The irregular line represents the subject's attempt to place his fundamental frequency between the two limits indicated by the horizontal lines.

To evaluate the degree of improvement in the intonation patterns of the subjects, a number of test phrases were recorded before and after the sequence of instructional sessions. The corresponding pre- and post-instruction responses were then recorded in succession, but their order was randomized, so that listening judges could not tell whether a particular sample was from a pre-instruction or a post-instruction recording. The listeners judged which of the two samples were more satisfactory. Examples of some pairs on which the post-instructional response was judged to be better than pre-instructional response by a large number of judges will now be presented (tape). The numerical result obtained from three groups of listeners and nine subjects are listed in Table II. It is seen that with the exception of subjects 1 and 8, all show improvement for listener group 1 (most sophisticated listeners), i.e. the post-test response is judged to be better than their pre-test response. For other (less sophisticated) listening groups, less favorable results have been obtained.

During instruction, a number of interesting observations have been made: (1) All subjects became aware that they could change their pitch and all improved their intonation after mastering their rhythm patterns for the phrases. (2) Initially, all attempted to change their pitch by increasing vocal intensity; later most of them could raise their pitch without significantly changing vocal intensity. All experienced more difficulty for rising rather than falling inflections. (3) Because of the excessive effort of subject 1 to improve the intonation pattern, the subject's utterances deteriorated in other ways so that the overall effect was judged to be inferior, while the intonation itself had obviously improved.

V. Conclusion

In view of the experimental evidence presented above, it may be concluded that with very few exceptions the intonation patterns of the subjects improved significantly after a short training period. Thus the instantaneous pitch-period indicator and the associated visual display can be used in the future as a worthwhile aid in speech instruction of the deaf.

Acknowledgments

The author wishes to acknowledge the enthusiastic cooperation of his research and consulting staff which made the presentation of this paper possible, namely Dr. W. Pronovost and Dr. D. C. Anderson of Boston University, and Mrs. S. Bass and Mr. N. Phillips of Northeastern University. This research was sponsored by the Vocational Rehabilitation Administration under Grant No. RD-2360.

References

1. Anderson, F., "An Experimental Pitch Indicator for Training Deaf Scholars," J. Acous. Soc. Am., 32, 1065-1074, 1960.
2. Dolanský, Ladislav, "An Instantaneous Pitch-Period Indicator," J. Acous. Soc. Amer., Vol. 27, No. 1, pp. 67-72, 1955.
3. Phillips, N. D., Teaching Vocal Pitch Patterns Using Visual Feedback from the Instantaneous Pitch-Period Indicator for Self-Monitoring, MS Thesis, Northeastern University, Boston, Mass., June, 1966.
4. Phillips, N. D., Remillard, W. J., Bass, S., Pronovost, W., "Teaching of Intonation to the Deaf by Visual Pattern Matching," American Annals of the Deaf, Vol. 113, No. 2, pp. 239-246, March, 1968.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-2-2

Automatic Detection of Intonational Signals in American English

Yukio Takefuta
Adjunct Assistant Professor and Research Associate,
Department of Speech, The Ohio State University, U.S.A.

Introduction

A graphic representation of fundamental frequency modulation in human speech plotted by using an instrument is often called a melody curve. Melody curves have been drawn for almost three-quarters of a century. Over these years, the instruments have been improved in many ways and today we can get a highly "accurate" representation of frequency modulation, even for modulation of very short duration. This representation, however, is not accurate in the sense that it represents the intonation of our language. The melody curve plotted by such instruments contains far more information than the human listeners possibly perceive as the intonational signals. If the intonation should be represented, transformation of the data and reduction of the redundant information are necessary and have to follow the "accurate" plotting of the melody curve.

Few researchers adopted a statistical method to analyze the melody curves and to predict the perception of intonational signals by human listeners. In this study, some basic statistical procedures were experimentally adopted to simulate the analysis of the melody curve possibly done by human listeners. This simulation of the perception of intonational signals by human listeners has been carried out by using the versatile capacity of a digital computer to automatically analyze

and store the data, and to make a decision according to predetermined instructions set by the programmer.

An intonational signal was defined to be any linguistically significant difference perceived by native listeners between the two utterances which are the same in lexical and syntactical structures. Accordingly, the algorithms for the automatic analysis of fundamental frequency were programmed so that the signal detection by the computer matched that of native listeners as closely as possible. No attempt was made to directly compare the graphic representations of fundamental frequency modulation or melody curves.

Experimental Procedure

Each of 10 male and 10 female speakers of American English recorded 40 English sentences. The recorded sentences which had the same lexical and syntactical structures were then randomly paired by the experimenter to get 400 stimulus pairs. A group of 16 college students listened to these 400 pairs. They judged with statistically significant agreement that 76 pairs out of the 400 contained intonational signals and 84 pairs did not. These 160 pairs (76 with intonational signal and 84 without) were in turn subjected to instrumental measurement of fundamental frequency with the Trans Pitchmeter. The values in Hz of fundamental frequency read every 40 msec. of the utterance were considered raw data of the stimulus which were fed to the computer.

The raw data of each utterance were first multiplied by a constant so that the difference in pitch of different speakers could be cancelled. The amount of deviation of the data from the mean was also adjusted, sentence by sentence, by multiplying the same constant to the difference of each value from the mean.

Then, assuming the psychological dimensions of intonational signal to be the pitch pattern, the pitch level, the pitch range and the duration, the corresponding physical dimensions were sought. The standard deviation of the adjusted raw data was calculated to represent the physical dimension of pitch range for each utterance. The pattern of data distribution was too complex to analyze when the set of data for the entire utterance was considered as a unit. In order to simplify

the physical representation of this dimension, the set of data for the entire utterance, arranged in order of time, was divided into subsets coinciding with the successive constituent "syllables" of equal duration. Each of these subsets was further subdivided into two halves to get subsets of "half-syllables." Each of the subsets of "syllables" or "half-syllables" thus obtained did not necessarily coincide with the actual length of syllables in the utterance. However, this division was convenient to simplify the representation of the pattern of data distribution. A linear regression line was fitted by a method of least squares to the raw data of each subset. The slopes of these regression lines collectively and in the order of occurrence were indicative of the pattern of distribution of data for the entire utterance. The frequency level to correspond pitch level was also obtained for each of these subsets by calculating an algebraic mean of the raw data within each subset. The mean values collectively indicated the variation of the level of fundamental frequency in the utterance. The physical counterpart of duration was obtained by dividing the total duration of vocalization by the number of subsets in each utterance.

Data for these four physical dimensions of intonational signals were stored into computer storage after a simple transformation of rescaling. Automatic signal detection in the intonations of paired utterances was carried out by comparing corresponding physical dimensions of the paired utterances and going through a decision tree, each bifurcation of which has a criterion value for the difference in each or combined data of physical dimension to be significant. The ordered set of rules which was referred to at each decision-making was as follows (the criterion values were empirically determined to be optimum for the particular transformation of the data in this study):

1. If the physical dimension of the frequency range is greater than 2.5 and less than or equal to 4.0, subtract 5 and 2 from the physical dimensions of the frequency pattern and the frequency level respectively.
2. If the constants multiplied to cancel the pitch difference of the paired speakers are both larger than 1.5 and the physical dimension of the frequency range is less than 1.5, add 5 and 2 to the physical dimensions of frequency pattern and the frequency range respectively.
3. If the total of three physical dimensions, the frequency pattern, the frequency level and the duration is larger than 16, intonational signal is present in the pair.

4. If the physical dimension of frequency pattern and that of frequency level are both greater than or equal to 5, intonational signal is present in the pair.

5. If the physical dimensions of frequency pattern, frequency level, and the duration are greater than or equal to 5, 4, and 5, respectively, intonational signal is present in the pair.

6. Otherwise, intonational signal is not present in the pair.

Results and Conclusions

All 160 pairs of utterances went through the simulated detection process. Seventy-one pairs were predicted to have had intonational signals and 89 pairs were predicted to have had no intonational signals between the pairs. When the prediction made by the computer was compared, pair-by-pair, with the criterion decision made by the group of native listeners, there were 15 errors. The percent of the correct prediction was 90.6. Six errors were made in 60 utterance pairs recorded by female speakers. The computer made errors in predicting the presence of intonational signal only twice in 51 pairs recorded by male speakers, while there were seven errors for 49 pairs recorded by a male and a female speaker. The percent of the correct prediction was 90.0, 96.1, 85.7 for speaker pairs of female-female, male-male, and male-female respectively. The percent correct decision of the native listeners varied from 95.6 to 76.8. The average accuracy was 87.5 percent. The detection of intonational signals by computer simulation was found to be as good as the detection by average human listeners.

The simulation of the human auditory perception in this study was made highly arbitrarily, so that all processes could be made automatic in spite of the limited data available. For example, there is no reason why the data on intensity should be excluded from this simulation, why the duration of subsets should be equally divided. When these apparent fallacies are remedied, the percent accuracy of the detection of intonational signals by computer may be raised to a satisfactory level for practical use. IBM 7094 spent less than half a second to analyze an utterance of two to three seconds by going through all the simulated detection processes described above.

This research was supported by a contract between The Ohio State University Research Foundation and the Office of Naval Research Nonr495(18).

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-2-3

Voice Aperiodicity in terms of Pulse Shape and Interval

Shinji Maeda

The University of Electro-Communications

Osamu Fujimura and Masayuki Sawashima

Faculty of Medicine, University of Tokyo

Introduction

Recent physical and physiological studies have shown that there is no complete periodicity in human voice and both the pulse shape and the time interval of the glottal wave fluctuate. Rigorously speaking, therefore, the fundamental period of voice must be defined for a particular method of the pitch detection employed.

Fig. 1 illustrates examples of the successive vowels of the fundamental periods, for a stretch of vowel utterance as determined by different methods. The pattern of fluctuation of the fundamental period markedly differs depending on the method of detection.

In this study, this point has been examined further by use of vowel synthesis and perceptual evaluations, to conclude which method best defines the pulse positions in a practical sense.

Experimental Procedures

Fig. 2 shows a block diagram of the vowel analysis and synthesis system. The speech and subglottal signals are recorded simultaneously. The speech signals are analized by a soundspectrograph and the formant frequencies

(F_1 ... F_5) and band widths (B_1 ... B_5) are estimated by visual inspection. These values fed to a computer are used to control the pole circuit of a terminal analog synthesizer.

The subglottal wave is low-passed with a cutoff frequency of 1 kHz and is sampled at a rate of 10,000/sec in 11-bit levels. First, the fundamental periods of the subglottal wave are determined by a peak detection method. In this method, the position of pulses of the glottal wave are defined as the time moments of the summits of salient peaks, and the fundamental period is defined as each time interval between consecutive pulses. Next, the pitch detection is performed by an autocorrelation method in reference to the results obtained by the peak detection. In Fig. 1, the curves (B) and (C) have been obtained by use of different time windows for the autocorrelation method.

Finally, in the "BUZZ SOURCE GENERATION," the buzz pulses are generated at the time moments determined by methods above. The buzz source signal is fed into a hardware terminal analog synthesizer. The pulse shape is triangular with fixed values of rise and fall times. The amplitude is also fixed except a gradual rise and fall at both ends of an utterance.

Results and Discussion

Informal listening test of the synthesized vowel samples have revealed that the determination of the pulse positions by the method (A) (peak detection) gives rise to a coarse voice, far from the normal original voice characteristics. The sample synthesized in accordance with the method (B) was fair in naturalness and the sample by (C) was the most natural.

The irregularity of the glottal wave of the natural voice must include variation of the pulse waveform from period to period. This influences the peak position, but the correlation coefficient between the consecutive pulse

may remain comparatively unaffected.

Consequently, the synthesized vowel could have excessively fluctuating periodicity as determined by the autocorrelation method, and probably by the human auditory mechanism, if the pulse shape was kept constant and still the pulse positions were determined by the peak position of the original voice signal. This seems to explain the experimental result that the synthesis by use of the peak detection method makes the voice characteristic abnormally rough.

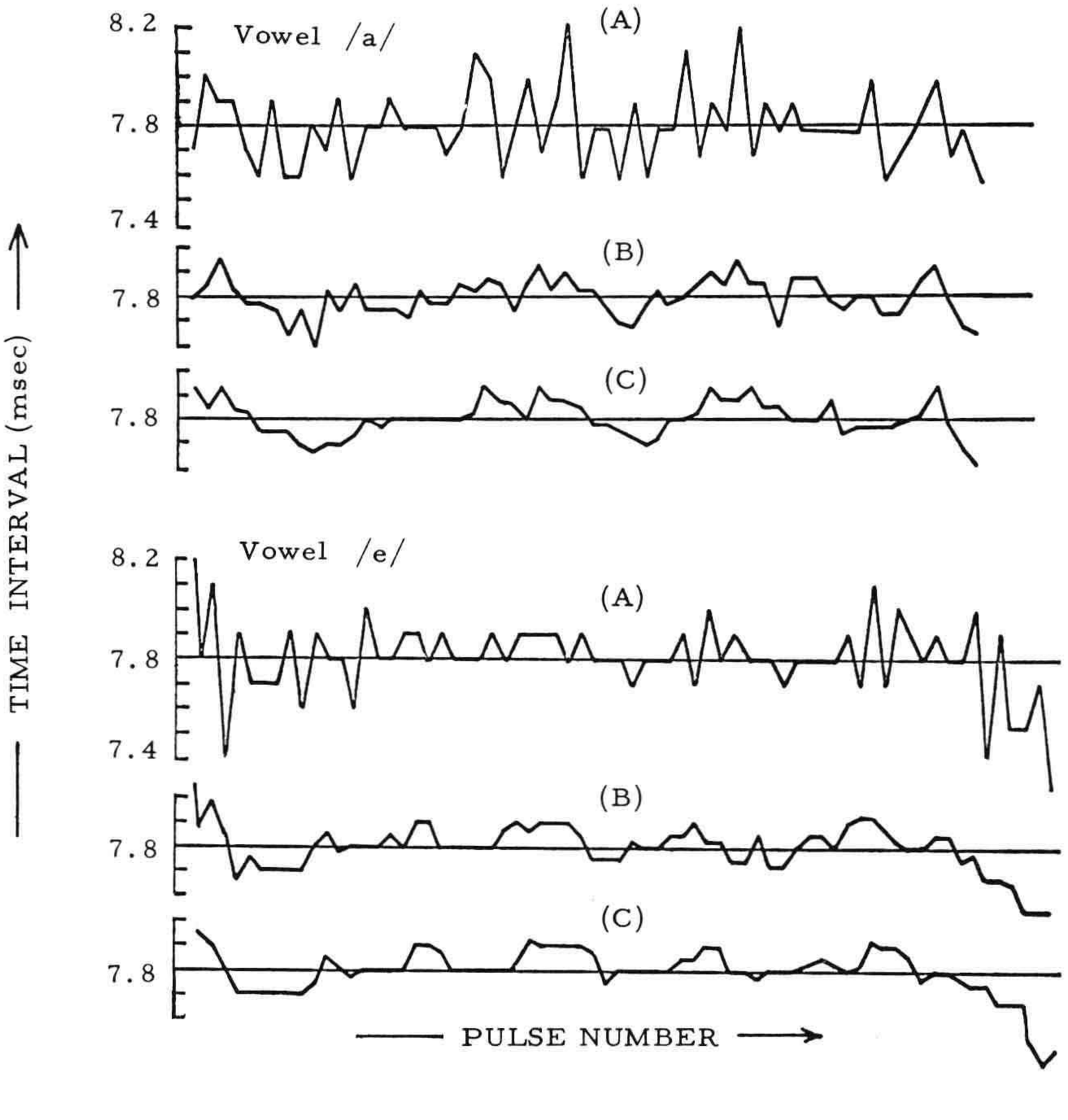


Fig.1 Time intervals between cosecutive pulses determined by (A) peak-detection method, (B) autocorrelation method using a ±3msec time window and (C) by autocorrelation with a ±8msec window.

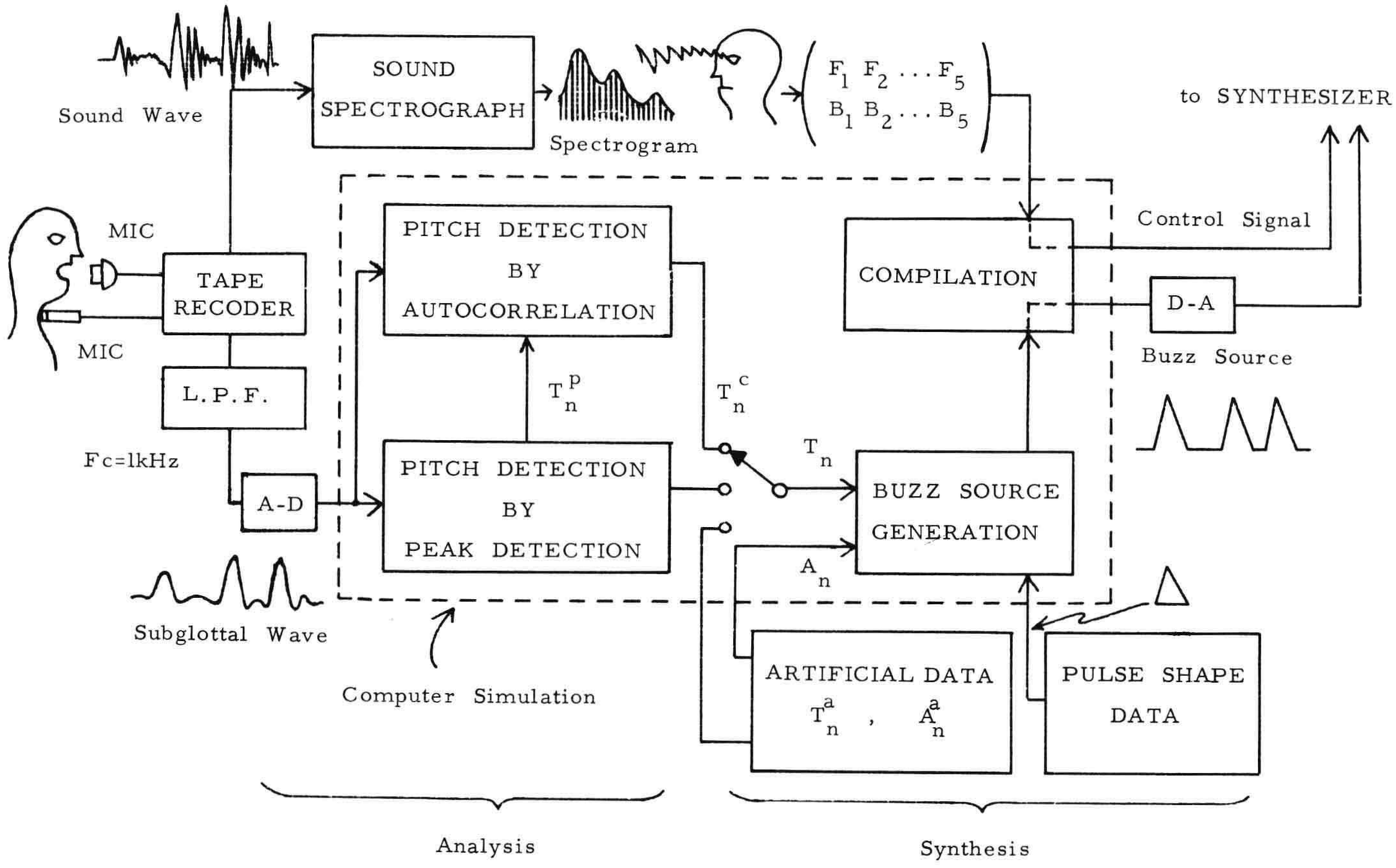


Fig. 2 Block diagram of vowel analysis and synthesis

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-2-4

Fast Formant Frequency Tracking Technique

Mamoru Nakatsui and Jouji Suzuki
The Radio Research Laboratories

This formant frequency tracking technique is an application of the inverse filtering method in the spectrum domain. Synthesized speech and two types of filter banks for spectrum analysis are used for evaluation of the proposed method. It is found that this technique can track the formant frequencies with errors less than one-half of the pitch frequency, and the time of calculation is one-hundredth to Analysis-by-Synthesis technique.

Principle

According to the generally accepted theory of speech production, the voiced speech sound is considered as an acoustic output of the vocal tract excited by the glottal source. If the vocal tract has no branch and the frequency range is limited (for example, below 3 kHz), the amplitude frequency spectrum envelope of the voiced speech is considered as composed of the following spectral elements: lower few formants, source, higher formants, and radiation. The later three are usually considered as a correcting term for the lower few formants. Thus, the spectrum envelope of the voice speech (H(f)) is

$$H(f) = \sum_{n=1}^{m} H_n(f, F_n) + KR(f) \qquad (1)$$

on a decibel scale, where $H_n(f,F_n)$ is the resonance character of the n-th formant whose frequency is F_n, KR(f) is the correcting term, and usually m=3 is adopted. Representatives of their relation are shown in Figs. 1-a and 1-b in the solid line.

Provided that the rough guesses of two of the first three formant frequencies are obtained in some manner that will be discussed in detail later, "simple resonance character" is extracted from the speech spectrum in the following manner. First, "quasi-speech spectrum" ($Q_i(f)$) omitting i-th formant character is synthesized as follows:

$$Q_i(f) = \sum_{n \neq i} H_{n0}(f, F_{n0}) + KR(f) \qquad (2)$$

where F_{n0} is guessed value for the n-th formant frequency. Subtracting Q_i from the speech spectrum, simple resonance character for the i-th formant (S_i (f)) is extracted as follows:

$$S_i(f) = H(f) - Q_i(f) \qquad (3)$$

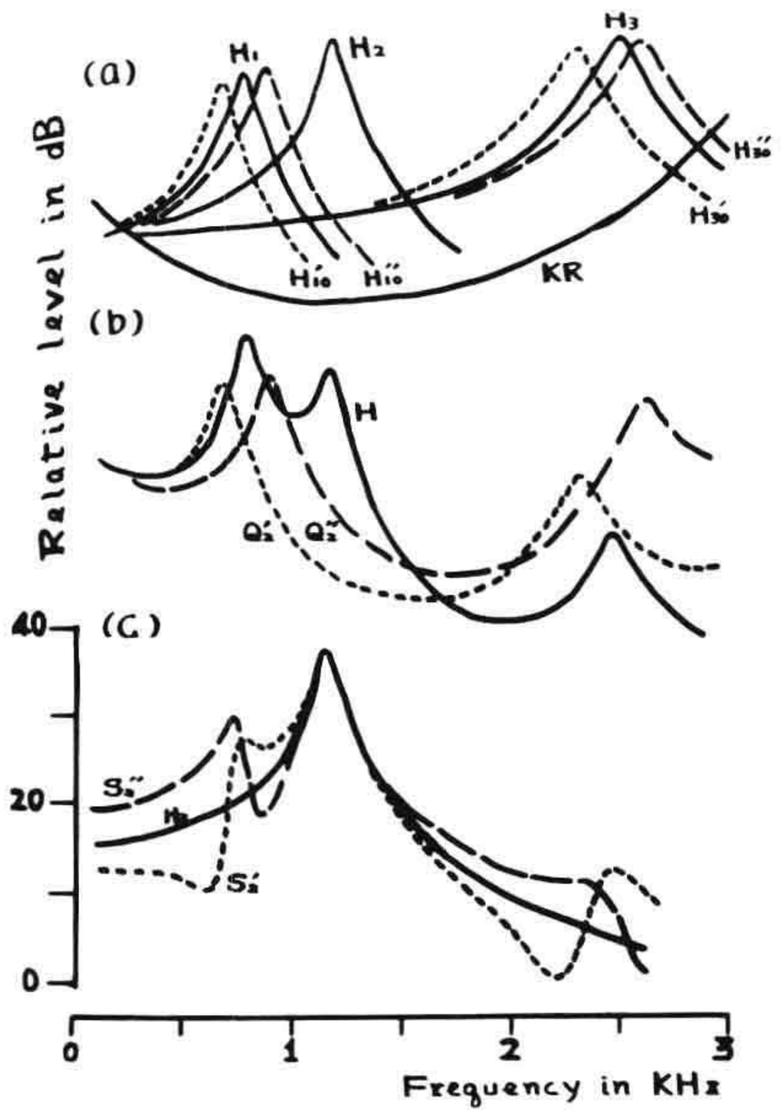


Fig. 1 Idealized illustration of the spectral composition of speech and extraction of the simple resonance curves. Solid lines show the input speech spectrum and broken lines the computed spectrum. (a) separate resonance curves and a correcting curve, (b) complete input speech spectrum and computed quasi-speech spectra, and (c) extracted simple resonance curves (second formant).

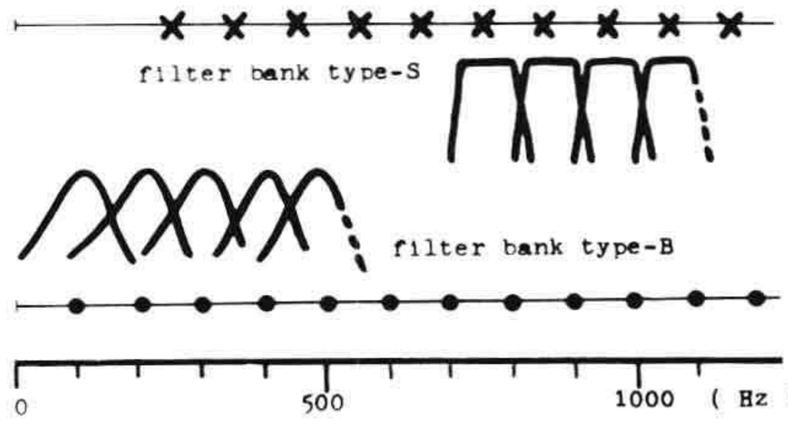


Fig. 2 Center frequency arrangement of the band-pass filters and the schematic illustration of pass band characters of the filter banks (lower channels)

As the representatives of their relations described in eqs. (2) and (3), quasi-speech spectra (the second formant omitted) and the extracted resonance characters for the second formant are shown in Figs. 1-b and 1-c respectivelly. In these figures $Hi0'$, Qi', and Si' represent the case where the guessed first and third formant frequencies are lower than those of input spectrum, and $Hi0"$, $Qi"$, and $Si"$ represent the higher case. Even though the guessed values are not so good, extracted resonance characters for the second formant exhibit the prominent peak related to its formant frequency.

Fairly good approximation for this formant frequency is obtained by calculating the first order moment arround it. In the same manner iterations are made of all three formants.

Analyzer and Synthesizing Speech

Discrete Fourier Transform or the filter bank is conventionally used as the means for measuring the short-time speech spectrum. Two types of filter banks are used. One (named type-B) is cascaded simple-tuned type which is simulated on the digital computer program and has equal frequency spacing and band-width in 100 Hz from 100 Hz up to 4000 Hz. The other (named type-S) is a real hard ware. The band-pass filters of this filter bank have sharp cut-off pass band character, and they are arranged so as to be 100 Hz step from 250 to 1150 Hz, 150 Hz step from 1275 to 2625 Hz, and 400 Hz step from 2900 to 4500 Hz. The frequency characters of both filter banks are sketched in Fig. 2. The outputs from them are rectified, smoothed, and sampled at the rate of 1/100 second. They are written as follows:

$$H(f_j) \qquad j = 1, 2, 3, \ldots\ldots, N \qquad (4)$$

where f_j is the center frequency of j-th band-pass filter and N is the total number of the band-pass filters.

In order to examine the performance of this procedure, artificial speech sounds are prepared. They are synthsized on the digital computer program in which five cascaded formant circuits are excited by the pulse train, and the first three formant frequencies are shown in Fig. 3 (solid lines). This mode pattern is associated with the movement of the formant frequencies in terms of the four tube approximation of the vocal tract in which the maximum constriction moves from glottis to lip position. It is considered as good representation of almost all the voiced sound situation. Fundamental frequency of the source takes four different values, 100, 120, 140, and 160 Hz.

Tracking Procedure

As described above, the first step of this procedure is the initial guess at the first three formant frequencies. Two kinds of guesses are made depending upon the tracking situation, that is, whether the spectrum is the first one or not within the same voiced interval. The resonance frequencies of the vocal tract having a uniform shape are used as the initial values at the first point of the voiced interval. There are used the sequencial constraint about the movement of the formant frequencies within the same voiced interval. Usually there is no discontinuity of formant frequencies between the adjacent spectra. Then the tracked final values of the formant frequencies are taken as the guesses at those of the next spectrum.

The first order moment (F_{i0}') is calculated about the extracted simple resonance character ($S_i(f_j)$) as follows:

$$F_{i0}' = \left(\sum_{j=i_1}^{i_2} h_i(f_j) f_j \right) / \left(\sum_{j=i_1}^{i_2} h_i(f_j) \right) \qquad (5)$$

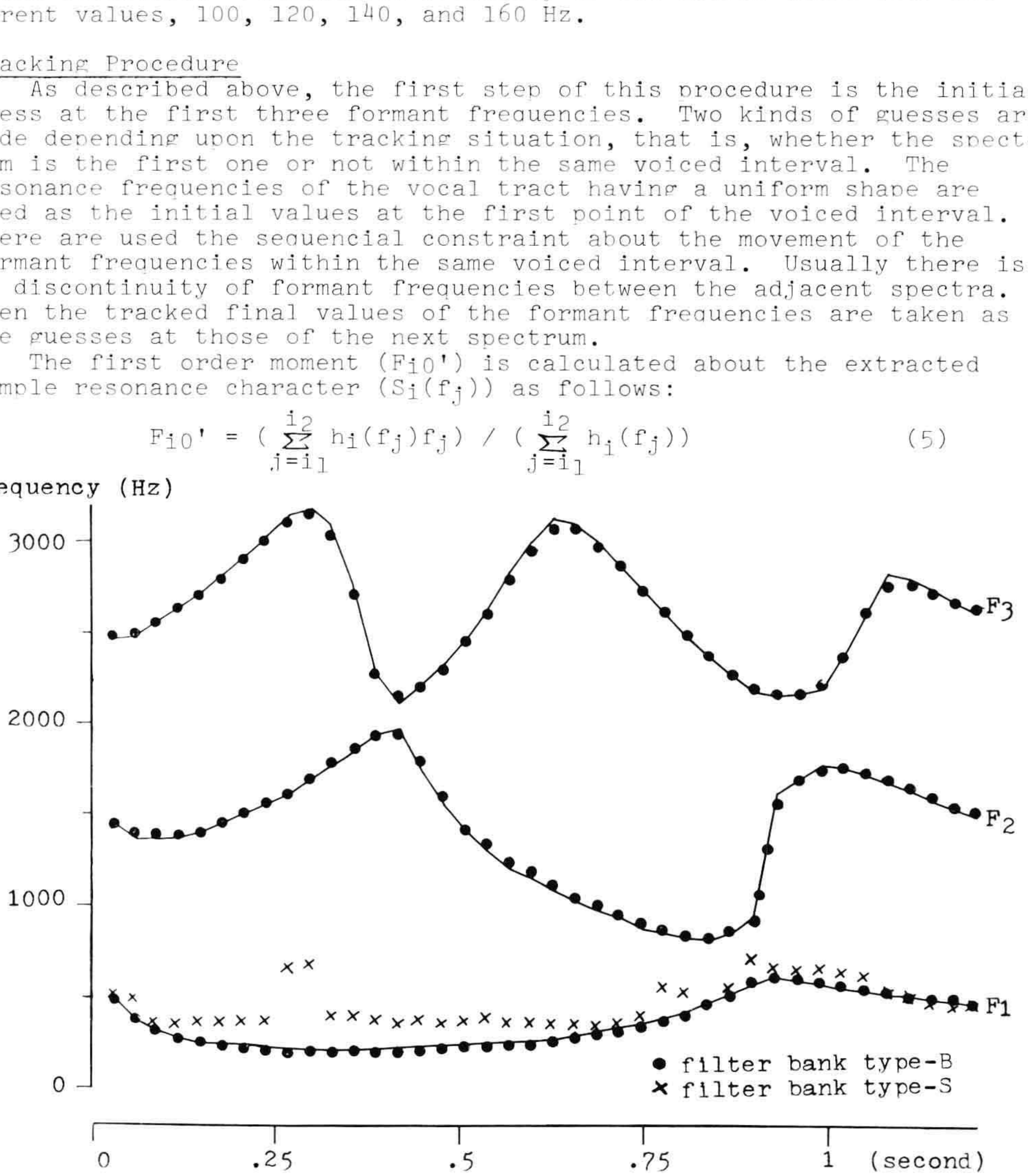


Fig. 3 Trajectories of the formant frequencies given in synthesis (solid lines) and the tracks of the formant frequencies.

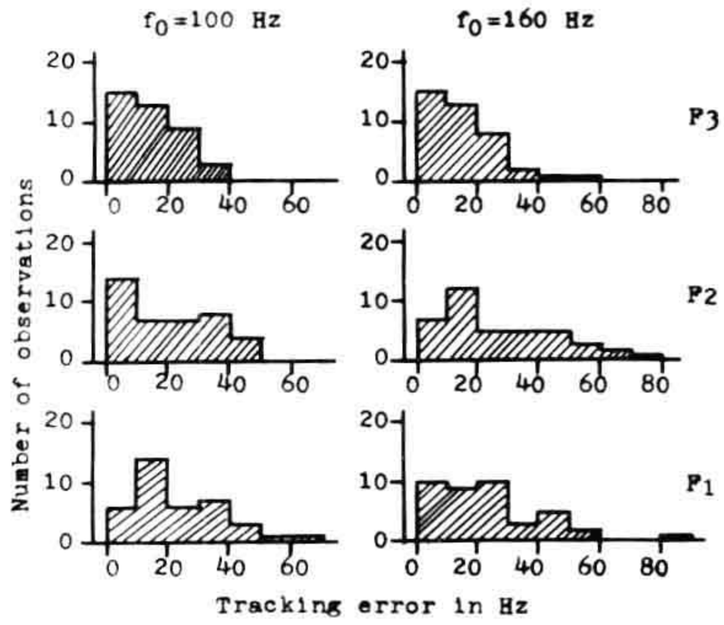


Fig. 4 Histgrams representing the distribution of tracking errors through the filter bank type-B. f_0 is the pitch frequency.

where $h_i(f_j)$ is the power description of $S_i(f_j)$, and i_1 and i_2 define the frequency range of the moment calculation. In the case of the initial spectrum, they are defined as follows:

$i_1 = 1$ (for F_1)
$i_1 = ch(F_{n-1}, 0)$ (for F_2 and F_3)
$i_2 = ch(F_{n+1}, 0)$ (for F_1 and F_2)
$i_2 = ch(3600)$ (for F_3)

In other cases they are defined as follows:

$i_1 = ch((1-0.15)F_{n0} - 200)$
$i_2 = ch((1+0.15)F_{n0} - 200)$

where ch(X) is the channel number of the band-pass filter which includes the frequency X within its pass band. The F_{i0}' just obtained is used for the synthesis of next Q_j ($i \neq j$).

In the same manner all three simple resonance characters are extracted and new guesses for these formant frequencies are obtained. These iterations are made until the differences between successively guessed formant frequencies are smaller than the previously given values which are chosen as 10 Hz to F_1 and F_2, and 20 Hz to F_3.

Result and Discussion

Typical results of the tracking experiment are shown in Fig. 3. Closed circles show the tracking results of the synthesized speech which has 100 Hz pitch frequency through the filter bank of type-B. Abrupt transition of the second formant and contiguities of the formants that are considered as the difficult conditions for the tracker like this are included in this synthesized sound. Their tracking results show good fits to the trajectories of the formant frequencies given in synthesis.

Crossed marks show the results through the filter bank type-S for the 160 Hz pitch. This tracking errors of the first formant are extremely large. Since the lowest channel of this filter bank is at 250 Hz in its center frequency and, further, in the left part of Fig. 3 the first formant frequency given in synthesis is nearly 200 Hz, it is understood that the output of the filter bank exhibits no prominent peak associated with this low first formant frequency. Duration of synthesized speech is 1.2 seconds and the average time spended in tracking is about 40 seconds.

The distributions of the tracking errors through the filter bank of type-B for the two pitch frequencies: 100 Hz and 160 Hz are graphed in Fig. 4. Their tracking errors may be within one-half of the pitch harmonics spacing. Tracking experiment in the natural speech is also made in Japanese vowels by 5 male adults and VCV type syllables by a male adult. Tracked formant frequencies of them show good fits to the formant bars in their spectrograms. The relations of the frequency spacing of the filter bank to the pitch harmonics and to the formant frequencies are generally complex and give some constraints on the performance of this tracker. Large errors in the F_1 tracking through the filter bank type-S is the worst of them. However, we might say that if filter bank such as type-B having fairly dense arrangement of filters along the frequency are used, we could expect precise tracking.

References

1. Fant,G.:Acoustic Theory of Speech Production, Mouton & Co., 1960
2. Kadokawa,Y. & K. Nakata:Formant Frequency Extraction by Analysis by Synthesis Technique, J. Radio Res. Labs., 10, p.251, 1963

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-2-5

A Simple Calculation of The Vocal Tract Configuration from Three Formant Frequencies

Yasuo Kadokawa and Jouji Suzuki
The Radio Research Laboratories

Introduction

The lossless vocal tract configuration is equally divided into six sections along its length. The necessary condition of the area functions in this case is obtained on the basis of the acoustical equivalent network theory. Giving the first three formant frequencies, the area function is represented by a polynomial with four degrees of the ratio of the cross-sectional area of the first section to one of the second section. Shroeder-Mermelstein's method is modified and applied so as to get the eighteen section area functions from this solution.

Degree of Freedom of the Constrained Area Functions

Uniqueness in the area functions is kept in the only case where the physiological constraints are relevant or both the first three formant frequencies and the first three antiformant frequencies are known[1]. At present it is very difficult to find the descriptive parameters on the articuratory movement in a dimension of the cross-sectional area through the X-ray observations and to obtain a number of formant and antiformant frequencies from the speech sound. But it is possible to extract automatically only the first three formant frequencies with high accuracy by the analysis-by-synthesis procedure. There is often used the physical vibration system of which the degree of freedom, though it is not limited at first, is reduced to the finite degree of freedom under the reasonable constraints. The same method is considered useful for the vocal tract system. Its application to the electrical network theory being taken into account, the acoustical equivalent network may be simpler in mathematical treatment than the wave equation[2].

The transmission matrix of the vocal tract is obtained by multiplication of each matrix in cascade connection from the glottis to the lips. Neglecting loss in the tube, the recurrence formula is given below:

$$(1)\quad K^N=\prod_{i=1}^{N}\begin{bmatrix}A_i & B_i\\ C_i & D_i\end{bmatrix}=\begin{cases}\begin{bmatrix}A^N & B^N\\ C^N & D^N\end{bmatrix}=\begin{bmatrix}A^{N-1} & B^{N-1}\\ C^{N-1} & D^{N-1}\end{bmatrix}\begin{bmatrix}A_N & B_N\\ C_N & D_N\end{bmatrix} & \text{FOR } \cos\theta_N=0,\ N\geq 2\\ \cos^N\theta_N\begin{bmatrix}a^N & b^N\\ c^N & d^N\end{bmatrix}=\cos^{N-1}\theta_N\begin{bmatrix}a^{N-1} & b^{N-1}\\ c^{N-1} & d^{N-1}\end{bmatrix}\begin{bmatrix}1 & b_N\\ c_N & 1\end{bmatrix}=\cos^{N-1}\theta_N\begin{bmatrix}a^{N-1}+b^{N-1}\cdot a_N\tan\theta_N & b^{N-1}+a^{N-1}\cdot\bar{a}_N\tan\theta_N\\ c^{N-1}+d^{N-1}\cdot a_N\tan\theta_N & d^{N-1}+c^{N-1}\cdot\bar{a}_N\tan\theta_N\end{bmatrix} & \text{FOR } \cos\theta_N\neq 0,\ N\geq 2\end{cases}$$

, where $a_N=jY_N=jA_N/\rho c_0$, $\bar{a}_N=jZ_N=j\rho c_0/A_N$, $a_N\bar{a}_N=-1$, $a_M\bar{a}_N=a_{MN}=-A_{MN}=-A_M/A_N$, $\theta_N=2\pi f l_t/c_0 N$,

N = total number of equally divided sections of the vocal tract,
l_t = tract length, A_i = cross-sectional area at i-th section,
ρ = density of air, c_0 = sound velocity.

In the case of voiced sounds, the transfer function is defined as the ratio of volume velocity at the lips to that at the glottis. Its formant frequency is obtained by either $D^N=0$ or $d^N=0$ on the assumption that glottal impedance is infinite and radiation effect at the lips is

neglected. There are the following two cases whether $\cos\theta_N$ is equal to zero or not. (1) $\cos\theta_N=0$ for $N\geq 2$. $D^N=d^N\cos N\theta_N$ and the lowest frequency derived from $\cos\theta_N=0$ gives the low-pass cutoff frequency of the i-th section. Their values are $f_c=c_oN/4l_t=1.0$, 1.5, 2.0, 2.5, 3.0, 3.5 kHz ...for N=2, 3, 4, 5, 6, 7... (2) $\cos\theta_N=0$ for $N\geq 2$. (a) $A_{MN}=1$. This means that all the cross-sectional areas take the same value, namely, the vocal tract configuration is a uniform tube. Substituting $a_Ma_N=-1$ into equation (1) and putting $d^N=0$, the n-th formant frequency is calculated by a simple equation such as $F_{on}=\theta_N(2n-1)=500(2n-1)$, where n=1, 2, 3 ... and $c_o/l_t=2\times10^3$. (b) $A_{MN}\neq 1$. This is the general representation of the vocal tract configuration. θ_N in equation (1) can be replaced by $\theta_{Nn}=2\pi F_{no}l_t/c_oN$ corresponding to the first three formant frequencies and they lead to the simultaneous equation that the area function is related to both the number of sections and the first three formant frequencies.

Constrained Area Functions with Six degrees of Freedom

If only the first three formant frequencies are known, it is impossible to secure x_{Nn} in terms of n-th formant above x_{N3}. Moreover, considered that the area functions are made more real by the appropriate interpolation, it is better to take six divided area functions. From equation (1), this polynomial is given by the following equation:

$$\begin{cases} p[1+q(1+r(1+t))]+q[1+r(1+s(1+t))]+r[1+s(1+t)]+t=\beta_1 & (2)_1\\ p[r(1+s(1+t))+t]+pq[s(1+t)+t]+pqrt+q[s(1+t)+t]+rt(q+1)=\beta_2 & (2)_2\\ prt=\beta_3, & (2)_3 \end{cases}$$

, where

$$\begin{cases} p=A_{12}=A_1/A_2,\ q=A_{23}=A_2/A_3,\ r=A_{34}=A_3/A_4,\ s=A_{45}=A_4/A_5,\ t=A_{56}=A_5/A_6, \\ \beta_1=(x_1x_2+x_2x_3+x_3x_1)/\Delta,\ \beta_2=(x_1+x_2+x_3)/\Delta,\ \beta_3=1/\Delta,\ \Delta=x_1x_2x_3 \\ x_n=\tan^2\theta_n=\tan^2(2\pi F_{no}l_t/6c_o) \quad \text{FOR } n=1,2,3. \end{cases} \tag{3}$$

In order to get the unique solution of the above equation, Shroeder-Mermelstein's method is deduced[1] and p=t, q=s are assumed. Substituting these relations into each equation $(2)_1$, $(2)_2$, and $(2)_3$, they can be rewritten in the forms

$$\begin{cases} r(1+p)^2q^2+2(1+p)(1+r)\cdot q+(2p+r)=\beta_1 & (4)_1\\ (1+p)^2q^2+2p(1+p)(1+r)\cdot q+p(2p+r)=\beta_2 & (4)_2\\ r=\beta_3/p^2. & (4)_3 \end{cases}$$

The above equation $(4)_1$ is settled on a polynomial with four degrees only p,

$$\begin{cases} [(\beta_1-\beta_2)^2-4(\beta_2+2\beta_3)]\cdot p^4+8\beta_3(1+\beta_1+\beta_2+\beta_3)\cdot p^3-2\beta_3[\beta_1(1+\beta_2+4\beta_3)+\beta_2(4+\beta_3)]\cdot p^2 \\ \quad +8\beta_3^2(1+\beta_1+\beta_2+\beta_3)\cdot p+\beta_3^2[1+\beta_2(\beta_2-2)-4\beta_3(2+\beta_1)]=0, & (5)_1\\ q=[-2\cdot p^4+(\beta_1+\beta_3)\cdot p^2-\beta_3(1+\beta_2)\cdot p+2\beta_3^2]/2[p^4+(1-\beta_3)\cdot p^2+\beta_3(1-\beta_3)\cdot p-\beta_3^2], & (5)_2\\ r=\beta_3/p^2. & (5)_3 \end{cases}$$

$$A_2=A_1/p,\ A_3=A_1/pq,\ A_4=A_1/pqr,\ A_5=A_1/pqr,\ A_6=A_1/pqr. \tag{6}$$

Arithmetic operations for the area function are done as follows: First, presenting the first three formant frequencies F_{no} and the tract length l_t, β_n is calculated from equation (3). Secondly, these values are substituted into equation $(5)_1$, then the positive p is obtained by equation $(5)_1$. Also q, r of equations $(5)_2$, $(5)_3$ are obtained from this value, respectively. Lastly, the cross-sectional areas from A_2 to A_6 normalized by the first one A_1 are derived from equation (6).

Area Functions Smoothed by Six Cross-Sectional Areas

The calculated area functions consist of only six sections, and their rectangular forms are far from the actual vocal tract configuration. In order to get more natural area function, the smoothing method by appropriate interpolation is considered necessary. After full discussion, Shroeder-Mermelstein's method[1] was modified by the exponential cosine functions with both odd and even terms so as to be fitted for six cross-sectional areas A_N (where N=1, 2, 3, 4, 5, 6). This smoothing function is written as follows:

$$A(x)=\exp.\left\{\sum_{N=1}^{6} m_N\cos[(N-1)\pi x/l_t]\right\}. \tag{7}$$

As A_N is known, the weighting function m_N is calculated from

(8) $A_N=A\{x=(2N-1)l_t/12\} = \exp.\{\sum_{N=1}^{6} m_N \cos[(N-1)(2N-1)\pi/12]\}$.

It is sufficient to divide the tract length into eighteen sections at the equal distance in the frequency range interested. The eighteen cross-sectional areas are obtained from the following equation:

(9) $A(x)=A\{x=(2N-1)l_t/36\} = \exp.\{\sum_{N=1}^{6} m_N \cos[(N-1)(2N-1)\pi/36]\}$.

Application and Discussion

In order to verify the applicability of this method of calculation to the real vocal tract configurations, test is made of quantized X-ray data by Chiba and Kajiyama[3], Fant[4], and Heinz, in five Japanese vowels, six Russian vowels, and six English vowels, respectively. At first, these area functions are copied exactly, and the corresponding first three formant frequencies are determined by analyzing the acoustical equivalent network with the equal interval h=0.5 cm. Next, the smoothed area functions satisfying these values are calculated and compared with the original X-ray data as shown in Figs. 1(a) to (d), where FNO and FN denote the original and the smoothed formant frequencies, respectively. F6R the formant frequency with respect to six divided sections with radiation effect, and FNR the smoothed formant frequency with radiation effect. The first cross-sectional area A_1 is chosen to fit the original X-ray data.

It remains an open question whether the solution of the algebric equation $(5)_1$ possesses only one or not under the condition that all the area functions should be positive. As the result of close investigation of all vowels, we have always only one reasonable solution.

Each of six divided tubes acts as a low-pass filter with the cutoff frequency 3 kHz. Non-smoothed area function having a formant frequency above this value, for example, /i/ in Fig. 1(b), can not match the original X-ray data. But it keeps exact fit below this value.

Fig. 1(a) shows that the best approximation to the smoothed area functions and to the corresponding formant frequencies is calculated with a neutral vowel /e/ whose area function is slightly deviated from the uniform tube. Figs. 1(a) to (d) show the tendency that the more the maximum constriction of articulators moves forward or the smaller it is, the larger is the difference between the original vocal tract configuration and the corresponding one obtained. The formant frequency and the vocal tract configuration which are obtained hold similar tendency. It seems to depend on the formant pattern of the vocal tract configuration rather than the defect of this simple method of calculation. With respect to /u/ in Fig. 1(d) having two constrictions in a vocal tract, it is very difficult to be fitted for the area function, but the difference of the corresponding formant frequencies is not so much. When the three formant frequencies by Mermelstein for six Russian vowels are given to this simple method of calculation, both the area functions obtained and the first three formant frequencies are nearly equal to his data[1]. This simple method of calculation in general gives a little higher values to the formant frequencies, especially the third formant for most vowels, but the calculated area functions fit well for the original one except /u/.

To investigate the practicability of this simple method of calculation the area functions at the middle point of the phoneme boundary for Japanese connected speech /era/ are computed and the result is exemplified in Fig. (2). In order to extract the first three formant frequencies with high accuracy from the spectrum analyzer, the analysis-by-synthesis procedure is followed. The extracted first three formant frequencies FNO' always contain radiation effect which makes FNO, the formant frequencies of the vocal tract itself, lower by approximately $8r_l/3\pi l_t$, where r_l is the lip radius[4]. Therefore, FNO is given by the equation $FNO \backsim (1+8r_l/3\pi l_t)FNO'$ and are applied as the input of this simple

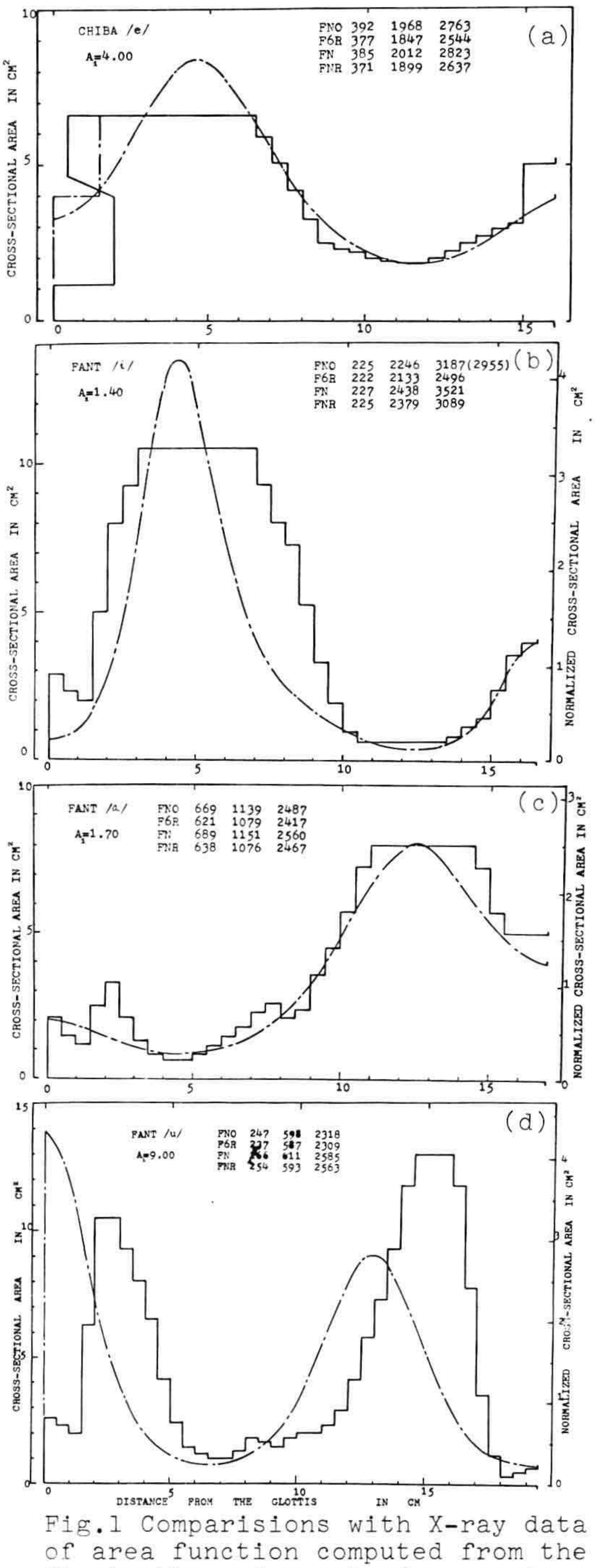


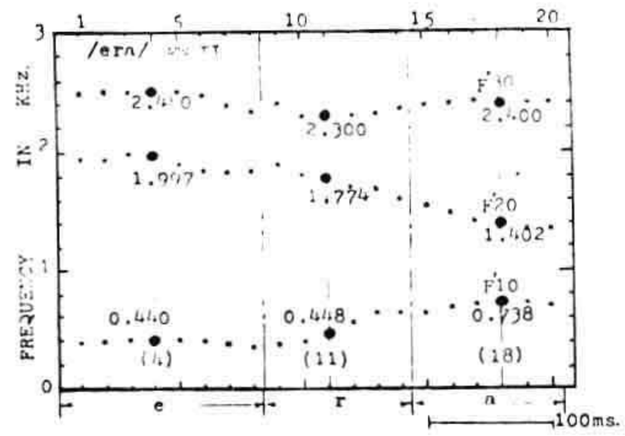
Fig.1 Comparisions with X-ray data of area function computed from the first three formant frequencies.

method of calculation. Although the X-ray data are lacking at the present time, the area functions obtained seem to give representation for the inherent shape of each phoneme.

In order to calculate the cross-sectional areas of the vocal tract from only the first three formant frequencies as simply as possible, the necessary condition of six equally divided sections was first obtained. Secondly, they were smoothed by the use of the result of Shroeder-Mermelstein's method. The results of quantitative analysis support the feasibility of articulatory description of the speech sound.

References

(1)P.Mermelstein, M.R.Shroeder, JASA vol.41 No.5 (May 1967).
(2)Y.Kadokawa et al., Journal of R.R.L. vol.11 No.54 (March 1964).
(3)Chiba and Kajiyama, The Vowel-Its Nature and Structure, Kaiseikan (1941).
(4)G.Fant, Acoustic Theory of Speech Production, Mouton and Co.,(1960).



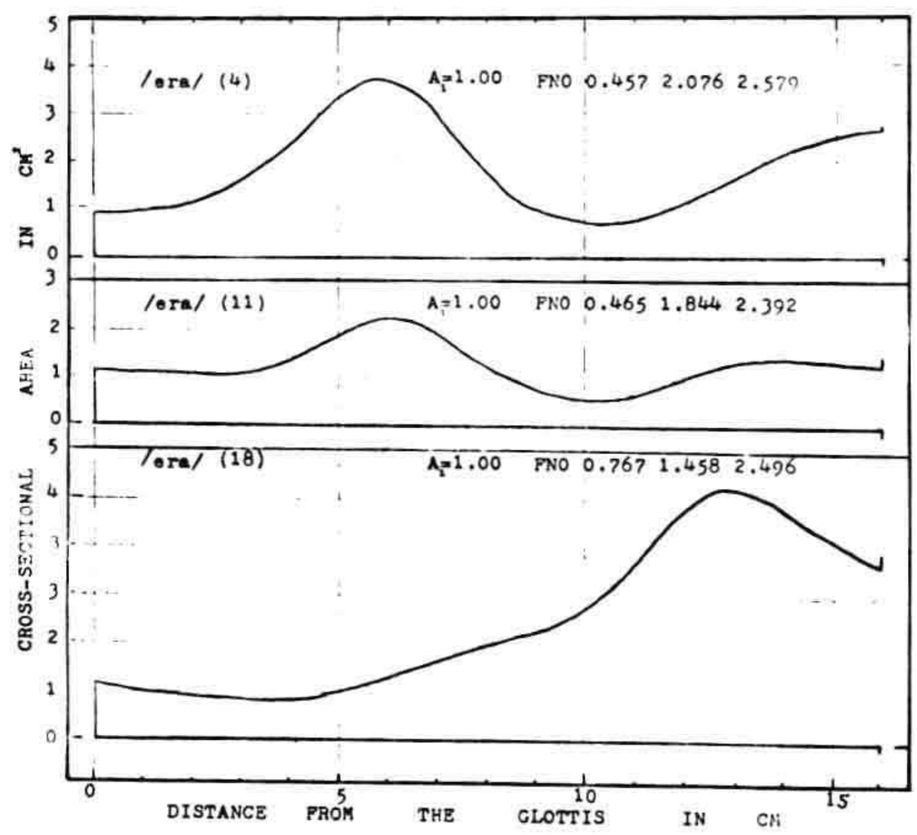


Fig.2 The first three formant frequencies extracted from A-b-S procedure (upper) and the area functions computed from these values (lower).

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-2-6

On the Formant Information in Connected Speech

Yoshinari Kanamori, Hiroyuki Kawai, Shizuo Hiki and Juro Oizumi

Research Institute of Electrical Communication,
Tohoku University, Sendai, Japan

INTRODUCTION

The purpose of this study is to synthesize the connected speech with the control signals of a minimum information rate. The rate is to be obtained from the perceptual contributions of each of the control signals to the quality of synthetic speech.

In this study, the control signals are limited to typical ones necessary to synthesize the phonemic feature. For this purpose, the characteristics of the formant are analyzed in the connected speech spoken in different tempos, and the information rate of the formant is calculated with these results and the difference limen (DL) of the formant frequency.

EXPERIMENTAL PROCEDURES

The speech materials used here are a weather forecast spoken in three different tempos by a male announcer whose dialect is typical of Tokyo Japanese; normally (Material N), slowly (Material S) and fast (Material F). Three parts of Material N are used for computation of the band width necessary to reproduce the smooth curve of each formant frequency with an accuracy of DL. These parts are named M. 1, M. 2 and M. 3, whose durations are 2.07, 1.45 and 1.68 sec.

respectively. The smooth curves are drawn along the center frequencies of the formant bars on the sound spectrogram, and in the silent and unvoiced intervals the curves are interpolated continuously. They are sampled every 7.8 msec. and quantized in about 9.2 Hz with the curve reader.

RESULTS AND DISCUSSION

Vowel Areas and Density Functions of Formant F1, F2 and F3.

Vowel areas measured from Material S and F are shown in Fig. 1. The vowels $\overset{1,2}{}$ are neutralized by means of contexts, and the positions of isolated vowels (marked X) remain in the edge of the vowel areas. It is only in special contexts that the formant frequencies are outside of the area surrounded by these points. The faster the tempo, the more the vowels concentrate toward the neutral vowel position. The arrows are drawn from X to the mean value of the formant frequency of vowel in connected speech.

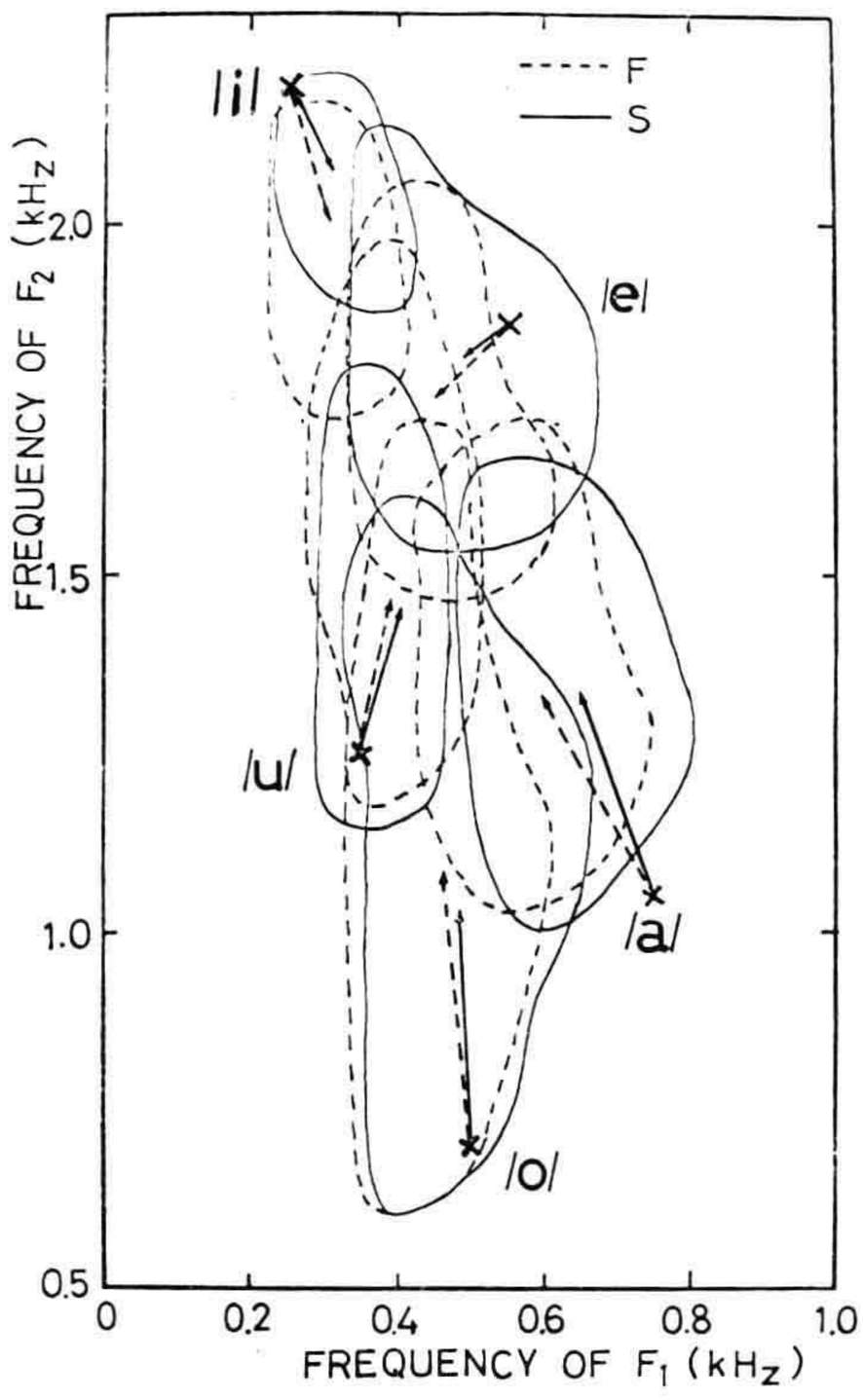


Fig. 1 Vowel areas measured from Material S and F

The density functions of the first (F1), second (F2) and third (F3) formants are shown in Fig. 2. Because of the inclination of occurrence frequency of vowels in the material, a remarkable peak (1.3 kHz) corresponding to vowel /a/ appears in the density function of F2. The density function of F3 is approximated by the normal density function. The neutralization by reason of the effect of tempo affects especially the density function of F3 and the mean value is decreased in frequency by about 160 Hz.

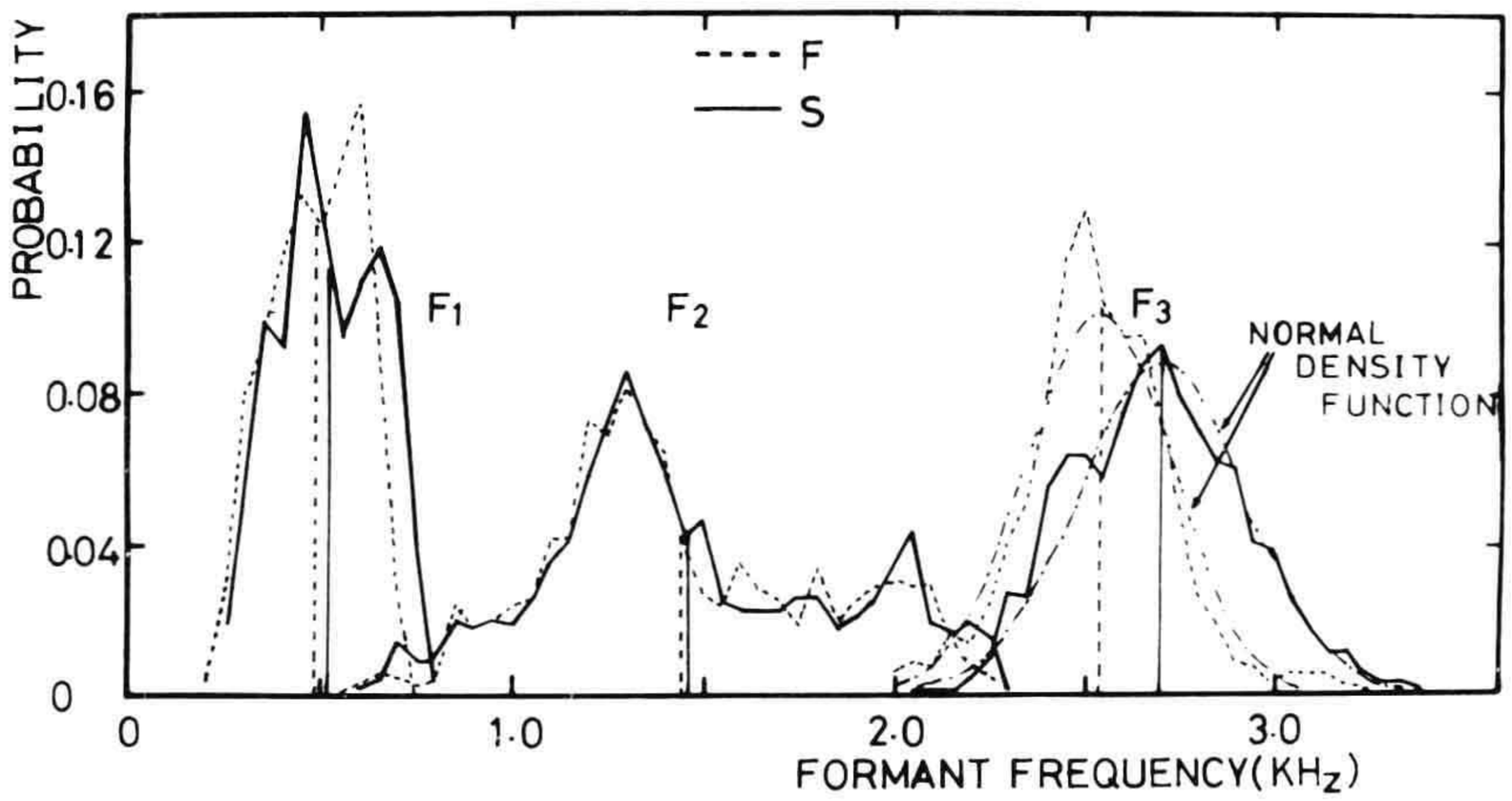


Fig. 2 Density functions of the first (F1), second (F2) and third (F3) formants measured from Material S and F.

Band Widths Necessary to Reproduce the Smooth Curves of Formant Frequencies with an Accuracy of DL.

This problem is described in literature(3), but a number of questions remain: the method of interpolating the silent and unvoiced intervals, the errors in the process of computation and the effect of the difference of the language (English vs. Japanese). The free hand and linear interpolations are used. The errors occur mainly in the process of coding the smooth curves. Since the error of about 3% is inevitable, DL of formant frequency used here is $\pm 5\%$. The band widths necessary to reproduce the smooth curves only in voiced intervals with an accuracy of $\pm 5\%$ are obtained.

Table 1 shows the results.

	M.1		M.2	M.3	MEAN
(Hz)	L	F	L	L	L
F1	20.1	22.0	11.4	21.0	17.5
F2	8.0	6.9	7.0	7.0	7.3
F3	8.6	9.0	7.0	6.0	7.2

M. 1, M. 2 and M. 3 are described in this paper.
F: Free hand interpolation
L: Linear interpolation

Table 1. Band widths necessary to reproduce the smooth curves of formant frequencies with an accuracy of DL. Tempo is normal (Material N).

CONCLUSION

The quantizing level of the smooth curve is determined according to DL. Since the contribution of F2 to the quality is much more significant than that of F1 and F3, the optimum sampling frequency is twice the band width of F2.

The results of Material N is used in the calculation of the information rate of formant frequency.

1) On the assumption that F1, F2 and F3 are independent of one another.

$$I(F1, F2, F3) = \sum_{i=1}^{3} \left\{ -\sum_{Fi} Pr(fi) \log Pr(fi) \right\}$$

$$= 3.58 + 3.26 + 1.59 = 8.43 \text{ bits/Hz}$$

Information rate IF = 8.43 x 14.6 ≒ 123 bits/sec.

2) On the assumption that F1 and F2 are independent of F3.

$$I(F1, F2, F3) = -\sum_{F1F2} Pr(f1, f2) \log Pr(f1, f2) - \sum_{F3} Pr(f3) \log Pr(f3)$$

$$= 5.88 + 1.59 = 7.47 \text{ bits/ Hz}$$

Information rate IF = 7.47 x 14.6 ≒ 109 bits/sec.

These results indicate a criterion of the formant information rate necessary for synthesizing a more general connected speech.

Reference

1 J. N. Sheare and J. N. Holmes, "An Experimental Study of the Classification of Sounds in Continuous Speech According to their Distribution in the Formant 1-Formant 2 plane",
in Proc. 4th. Int. Cong. Phonetic Sci., Helsinki, 1961.

2 B. Lindblom, " Spectrographic Study of Vowel Reduction",
J. Acoust. Soc. Am. 35, 1773-1781, 1963.

3 J. L. Flanagan, " Band Width and Channel Capacity Necessary to Transmit the Formant Information of Speech",
J. Acoust. Soc. Am. 28, 139-142, 1956.

4 J. L. Flanagan, " A Difference limen for Vowel Formant Frequency",
J. Acoust. Soc. Am. 27, 613-617, 1955.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-2-7

Statistical Properties of Speech Sounds

Hiroya Fujisaki and Kunihiko Niwa
Faculty of Engineering
University of Tokyo, Tokyo, Japan
Masahiro Iwasaki, Akio Kameoka and Nobuyuki Goto
Central Research Laboratory
Tokyo Shibaura Electric Co., Kawasaki, Japan

For the realization of an efficient speech communication system, characteristics of the channel including modulation have to be matched to the statistical properties of the signal. Although there are important contributions already published in this regard, few of them relate these properties to the temporal structure of speech, which can not be considered as a simple ergodic signal, since it is composed of segments that differ vastly from each other in their statistical properties. Analysis has thus been made of the nonsense syllables conventionally used for the articulation test of Japanese, after a coarse segmentation based on the characteristics of excitation sources and vocal tract configurations.

Analysis Procedure

The speech material consisting of 100 nonsense syllables of Table 1 was recorded in an anechoic chamber by a male voice. Sonagrams of these syllables, as well as photographic records, were made prior to the analysis in order to provide means for segmentation. Although a finer segmentation was possible, it was considered appropriate for the present purpose to classify portions of each syllable by a rather coarse segmentation based on the following procedure.

In the first place, a steady vowel, defined by the segment in which the second formant frequency lies within 10 % of its stationary value, is separated from each syllable. Except for a few cases, the remaining portion can be further divided into two segments either by the presence/absence of the voice excitation or the character-

istics of the vocal tract transfer function. The first segment may be voiced or voiceless, while the second segment is always voiced. For instance, in the case of syllables starting with voiceless plosives, the first segment contains plosion, frication and aspiration, while in the case of syllables starting with nasal consonants, it consists of a voiced nasal murmur. The second segment, on the other hand, is always voiced and represents the transition of vocal tract transfer from the initial segment to the following steady vowel. According to this segmentation, the first segments are found to have the most contribution to syllable articulation. The total duration of the first segments in the 100 syllables, is 11,334 msec, of which 7,208 msec are voiceless, while the total duration of the second segments is 7,674 msec.

The signal was band-limited to 8 kHz upon playback, sampled at a rate of 16,000 samples/sec, quantized into 9 bits and was stored in a digital computer. Each segment was then analized to give both the instantaneous amplitude probability density function and the frequency spectrum.

Probability Density Function for the Instantaneous Amplitude

Figure 1 shows the average probability density function for the instantaneous amplitude of the first segments (voiceless (a) and voiced (b)), the second segments (c) and the steady vowel segments (d). It is seen that the experimental curves for segments (c) and (d) match the negative exponential function fairy well within the amplitude range of $\pm 3\sigma_v$, where σ_v is the root-mean-square value of steady vowel segments. This can be explained by a simple model of these segments assuming a periodic excitation of resonant systems. On the other hand, the first segments show characteristic density functions which are different from the negative exponential or the normal distribution. Figure 2 shows the cumulative amplitude probabilities, plotted on a normal probability paper, for voiceless first segments of /s/, /t$\int$/ and /t/ as well as the average, indicating that the normal distribution model applies approximately for /s/ and /t$\int$/, but not for /t/ and the average. Since voiceless fricatives can be regarded as the response of a linear system to a random noise excitation, a normal distribution gives a fair approximation. Voiceless plosives, on the other hand, contain plosions which can be regarded as the step response of a

linear system and therefore result in overall density functions which deviate from the normal model. The deviation is still greater in the case of voiced first segments, which contain both noise and buzz responses.

Frequency Spectrum

The average frequency spectra of the segments are shown in Fig. 3. Except for the frequency range from 3 kHz to 5 kHz, segments (c) and (d) are predominant and determines the overall spectrum. The spectral envelopes of these two groups are quite similar and fall off with a slope of -12dB/oct. at higher frequencies. On the other hand, spectra for segments (a) and (b) have considerably lower levels and contrast each other. Approximate expressions for the average amplitude spectra of these segments are,

$$S_a(f) = \frac{W_a}{\sqrt{[1 - (f/f_a)^2]^2 + [\frac{1}{Q}(f/f_a)]^2}} \qquad \text{for segment (a),}$$

$$S_b(f) = \frac{W_b}{\sqrt{1 + (f/f_b)^2}} \qquad \text{for segment (b),}$$

$$S_c(f) = \frac{W_c}{1 + (f/f_c)^2} \qquad \text{for segment (c),}$$

as illustrated in Fig. 4.

Conclusion

Analysis has been made of statistical properties of speech sounds, taking into account their temporal structure. The results have been correlated with the mechanism of speech production and simple models have been derived to serve as a basis for theoretical evaluation of speech communication systems in general.

Table 1. 100 nonsense syllables for articulation test of Japanese.

	Initial Consonants							
	None	Stops	Fricatives	Affricates	Nasals	Semi-vowels	Liquids	Palatalized Consonants
voiceless		pa ta ka	sa ʃa ha	tʃa				pja kja hja
		pi ki	ʃi hi	tʃi				
		pu ku	su ʃu hu	tsu tʃu				pju kju hju
		pe te ke	se he					
		po to ko	so ʃo ho	tʃo				pjo kjo hjo
voiced	a	ba da ga	za	dʒa	ma na	wa ja	ra	bja gja mja nja rja
	i	bi gi	zi		mi ni		ri	
	u	bu gu	zu	dʒu	mu nu	ju	ru	bju gju mju nju rju
	e	be de ge	ze		me ne		re	
	o	bo do go	zo	dʒo	mo no	jo	ro	bjo gjo mjo njo rjo

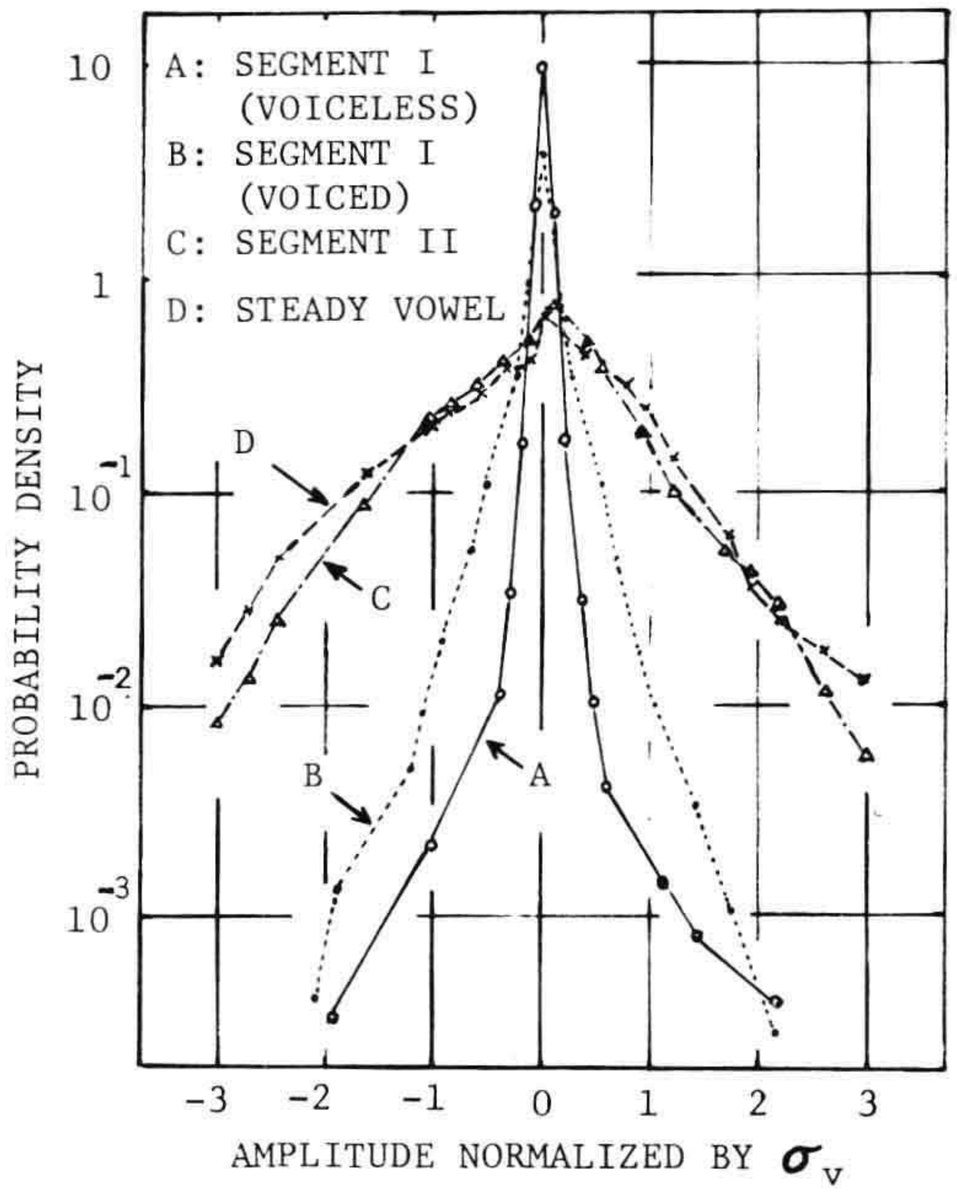


Fig. 1 Probability density functions of the instantaneous amplitude of speech segments.

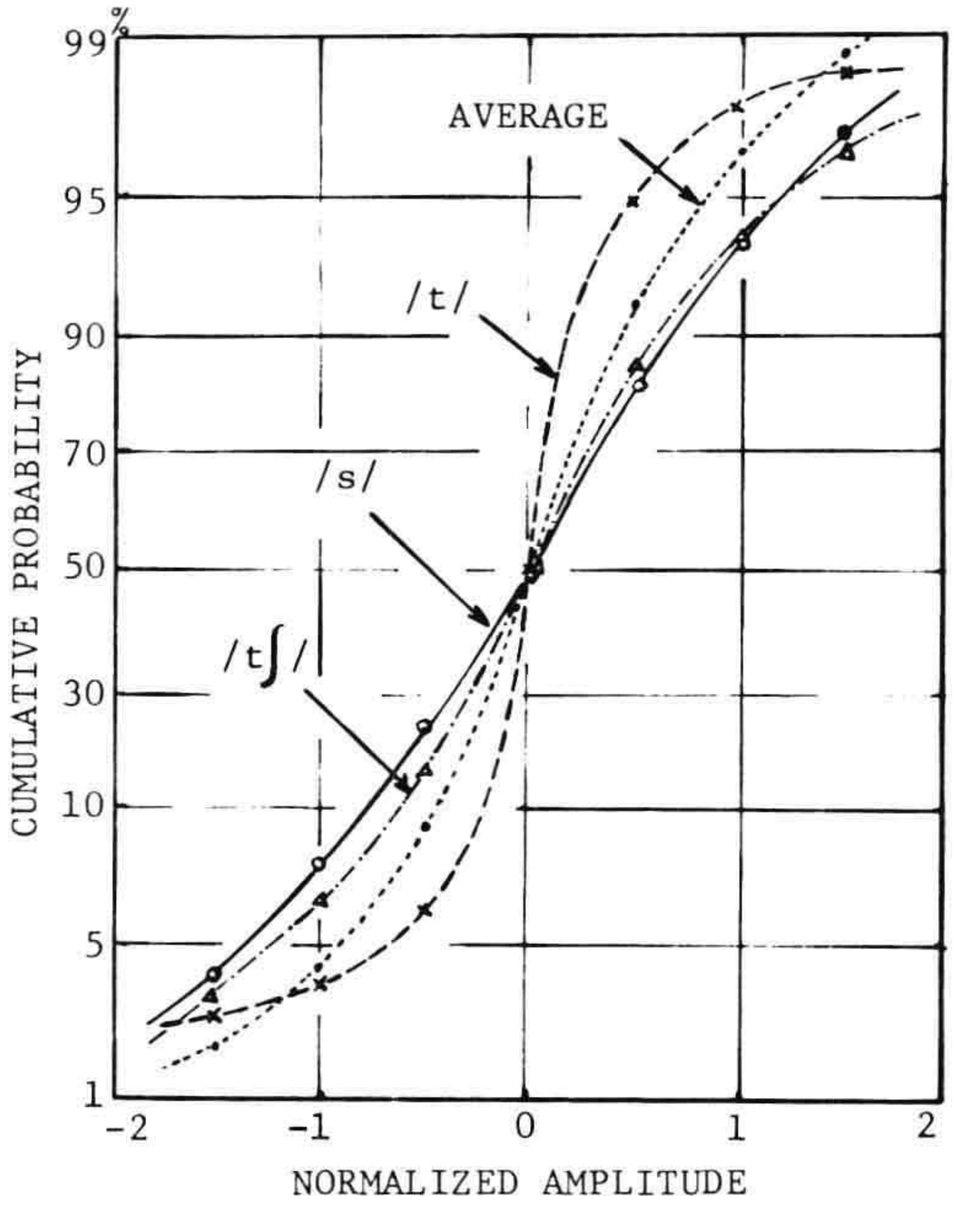


Fig. 2 Cumulative distributions of the instantaneous amplitude of voiceless first segments.

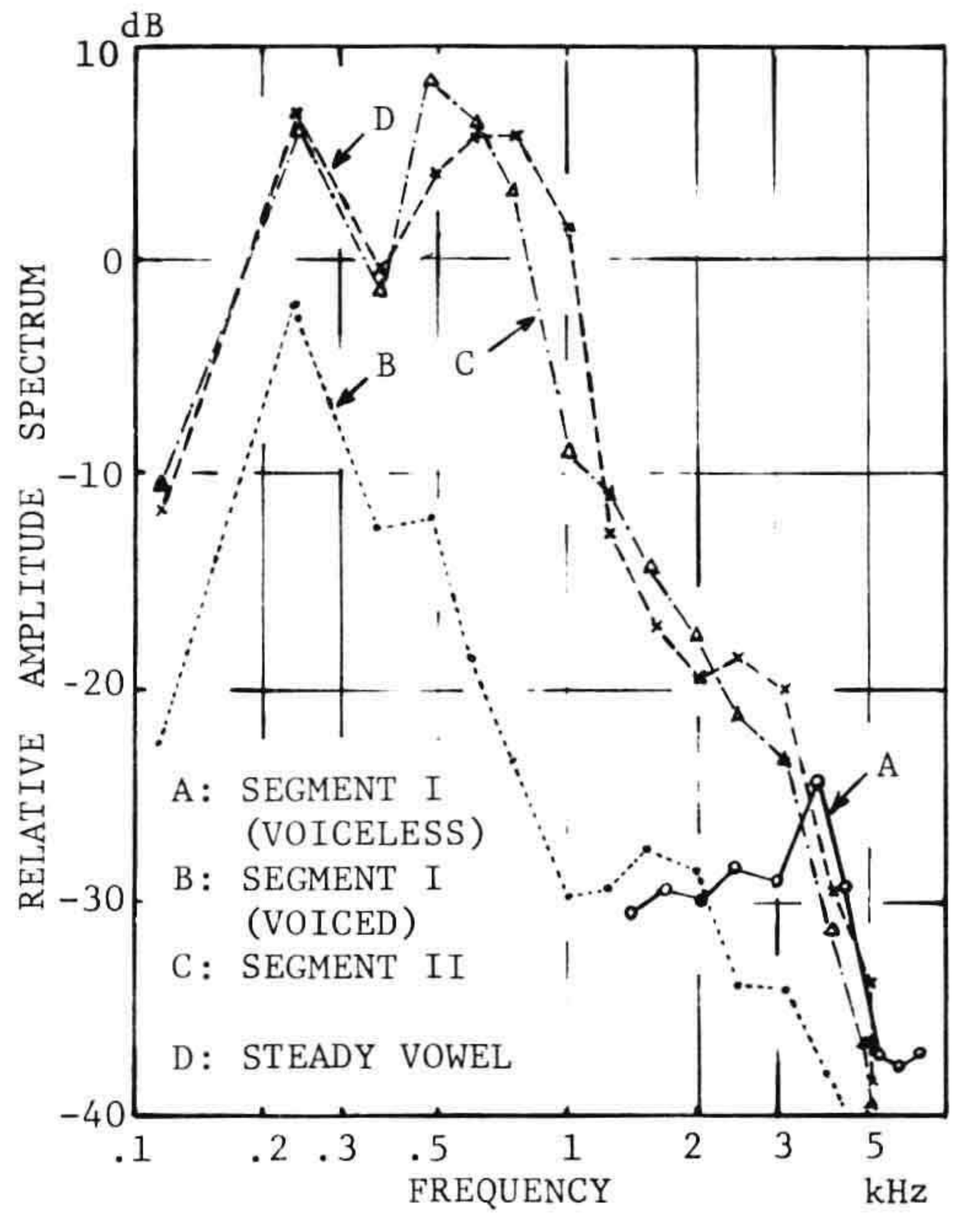


Fig. 3 Average amplitude spectra of speech samples.

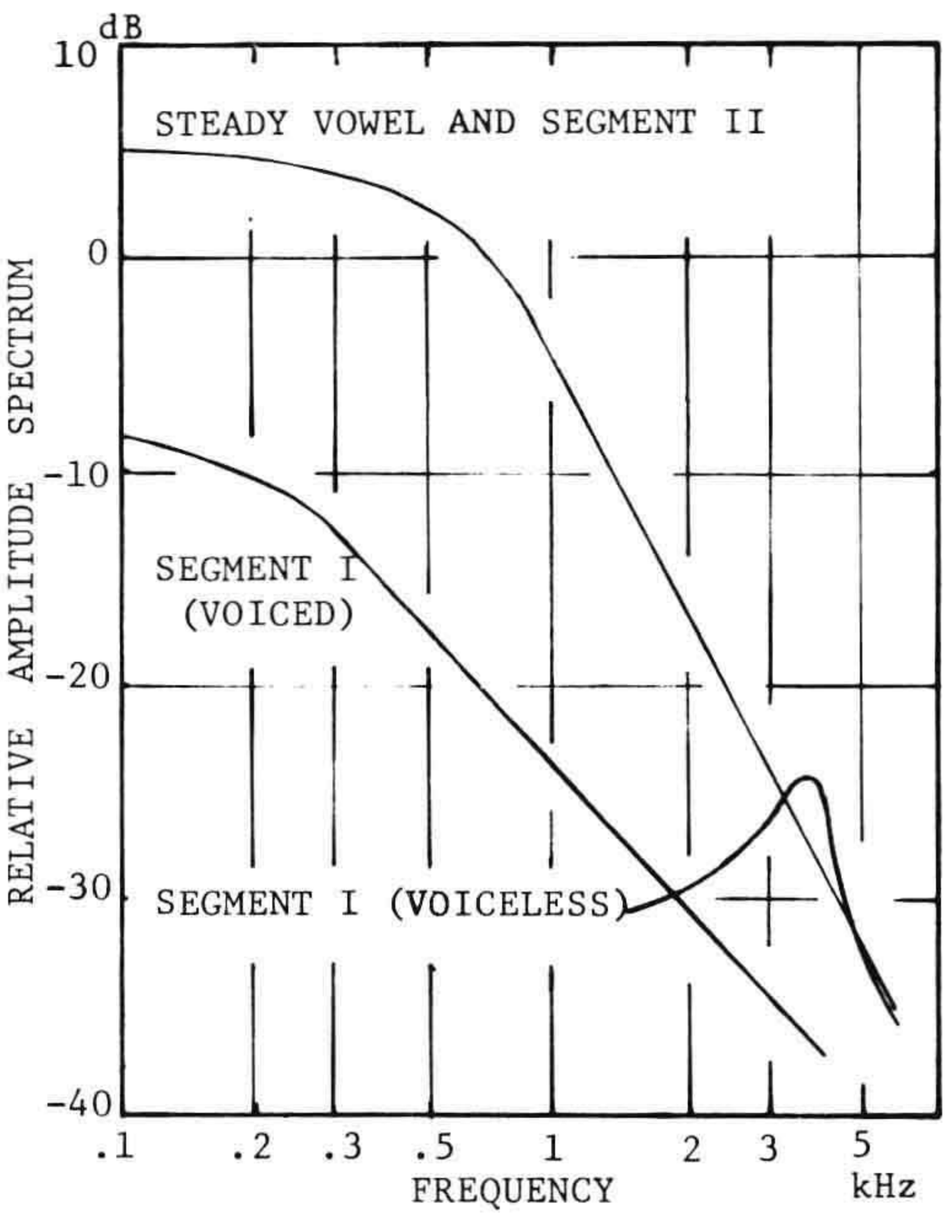


Fig. 4 Models for the average amplitude spectra of speech segments.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-2-8

Identifikation von menschlichem Verhalten durch statistische Analyse von Sprache

Identification of Human Behaviour by Statistical Analysis of Speech

Fritz Winckel und Manfred Krause
Institut für Kommunikationsforschung
Technische Universität Berlin

Ziel dieser Untersuchung ist, mittels eines von uns entwickelten Geräts zur Messung der Nulldurchgangsdichte von Sprachlauten bei variablen Schwellwertpegeln Merkmale für die Klassifizierung von Sprechern und von Einzellauten zu erhalten. Im ersten Fall ist das Gerät auf Langzeitanalyse, im anderen auf Kurzzeitanalyse geschaltet. Es war für die erste Anwendung von uns bereits im Bericht A 41 ICA Kongreß Liège 1965 beschrieben worden.

Das Meßgerät besteht aus einem Terzfiltersatz und einem Schwellwertanalysator, mit dem für jeden Filterbereich die Häufigkeit der Amplituden, die einen mittleren Lautpegel (z.B. 30 dB unter Vollaussteuerung) überschreiten, mittels Zähler gezählt wird. In der zitierten Arbeit wurde die aus der Nulldurchgangsanalyse ermittelte Impulshäufigkeit über die Terzfilterstufen aufgetragen, wobei die Impulszahl Z auf die Mittenfrequenz des jeweiligen Terzfilters f_m und die Gesamtsprechdauer t_{ges} bezogen wurde (fig.1). Dabei ergibt sich infolge natürlicher Abnahme der Amplituden der Partialtonfolge eine mit der Frequenz abnehmende Impulsdichte ($1/f_m$). Entsprechend einem auf -30 dB eingestellten Schwellwert nehmen die gemessenen Zählwerte gegen 4000 Hz zu Null ab. Deshalb wurde außerdem in einer zweiten Darstellung der Mittelwert der Impulszahl Z_m bezogen auf die Gesamtmeßzeit über den Terzfilter-Mittenfrequenzen aufgetragen (fig.1).

In diesem Fall wird der Abfall der Partialtöne durch den Anstieg mit ω kompensiert. Damit ist der Verlauf von F_2 und F_3 deutlicher ablesbar, der Frequenzbereich erweitert und der Übergang zum Konsonantbereich kenntlich.

Weiter wurde versucht, den Konturenkontrast der Diagramme zu verstärken, indem die Sprache mit einer EMT-Hallplatte künstlich verhallt wurde, wodurch die Zahl der Nulldurchgänge erhöht wurde. Fig.1 zeigt den Einfluß der Verhallung von 2 sec (Hallanteil gleich Direktschall, d.h.Hallradius).Eine eindeutige "Verstärkung"(Vermehrung der Nulldurchgänge im Mittel um das 2,2-fache) ist jedoch nur bis zu etwa 2000 Hz (Schwellwertpegel -30 dB) möglich, da bei höheren Frequenzen die Hallanteile wegen des Frequenzgangs der Nachhallzeit und des Übertragungsmaßes der Hallplatte abnehmen. Im unteren Bereich steigt die Zählrate nahezu proportional der Frequenz (gestrichelte Kurve)- fig.1.

Ein Kennzeichen der meisten statistischen Sprachspektren ist der Einbruch in der Häufigkeitskurve im Bereich von 160 bis 250 Hz für männliche Stimmen und etwa eine Oktave höher für weibliche Stimmen. Dies bedeutet, daß der Bereich unterhalb jenes Einbruchs der Tonhöhenverteilung F_o zuzuordnen ist, die sich, wie bekannt,als Gaußverteilung ergibt. Ein erstes Individualmerkmal ist die Streubreite dieser Verteilung, indem mit der wachsenden Tonhöhenvariation die Vitalität des Sprechers verknüpft ist. Die Streubreite kann so groß werden, daß der erwähnte Einbruch überbrückt wird und ein direkter Übergang des Gaußkurvenanstiegs in die Kurve des Formantbereichs erfolgt. Untersuchungen mit engeren Filterstufen, z.B. mit 6% relativer Bandbreite -statt des Terzfilters - brachten keine detaillierten Merkmale in den Impulsverteilungskurven. Die Terzteilung ist jedoch nicht ausreichend für eine genaue Ermittlung der mittleren Grundtonhöhe, und außerdem kann dadurch die symmetrische Gaußverteilung verzerrt werden, wogegen die

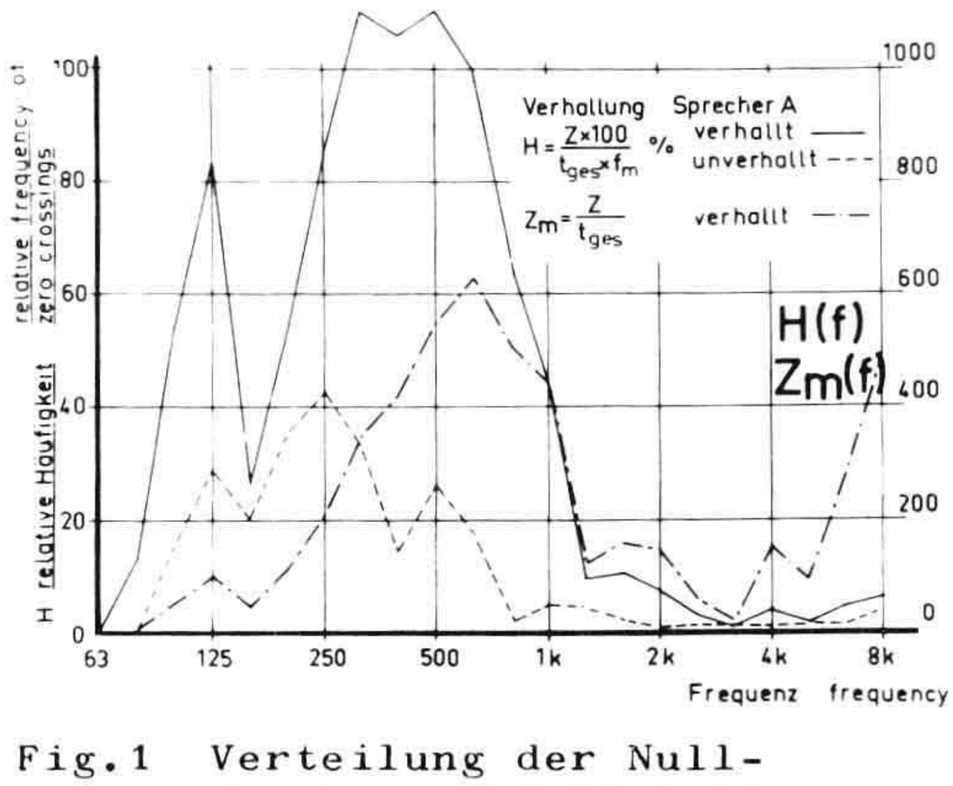


Fig.1 Verteilung der Nulldurchgangshäufigkeit
--- prozentual (Normalkurve)
___ prozentual mit Konturenverstärkung
-.- mittlere Zählrate mit Konturenverstärkung

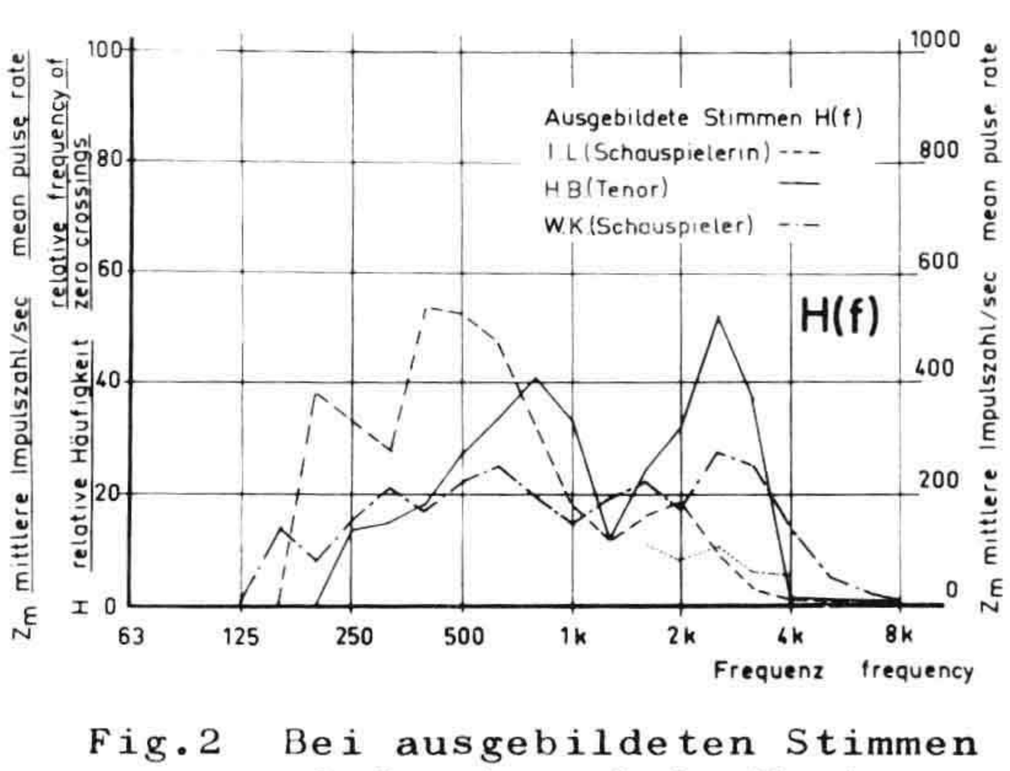


Fig.2 Bei ausgebildeten Stimmen sind extrem hohe Maxima im Bereich von 2 - 3,5kHz nachweisbar
(....unausgebildete Sprechstimme zum Vergleich)

6%-Filterteilung diese Verteilung genauer wiedergibt.

Die Zusammenhänge zwischen den statistisch wenig streuenden Abmessungen des Vokaltraktrohres von 17cm Länge bei männlichen Personen, dessen Eigenresonanz dem ersten Formanten F_1 entspricht, den zugehörigen Oberwellen und den Konstriktionen des Rohres sind physikalisch eindeutig und akustisch bestimmt durch die Vokalfarbe. Die Untersuchung zeigt, daß bei normalen Stimmen die statistische Häufigkeitsverteilung für den Bereich F_1 und F_2 nur geringe Streuungen der Maxima aufweist. Dagegen findet man individuell unterschiedliche Werte des F_1/F_2-Verhältnisses der Häufigkeit. In Fällen von Phonasthenie, die mit einer Muskelschwäche der Artikulationsmechanik und oft zugleich der Kehlkopfmuskulatur einhergehen, zeigen sich Abweichungen der Impulsdichten über den Terzbereichen gegenüber einer Normalkurve. Typisch ist ein relativ steiler kontinuierlicher Abfall über den ganzen Bereich F_1, was hörbar wird durch eine Verarmung des Vokalkolorits.

Im Frequenzgebiet F_3/F_1 können Individualmerkmale eher festgestellt werden. Während bei Normalstimmen der Mittelwert des Zählraten-Quotienten $\frac{Zm(F3)}{Zm(F1)}$ zwischen 0,4 und 3.0 beträgt, fällt er z.B.bei Phonasthenie auf 0.15 ab. Bei ausgebildeten Stimmen, wie Sängern und Schauspielern, erreicht er mit >5.0 wesentlich höhere Werte. Ein überragendes Maximum von F_3 in der Enveloppe des Häufigkeitsspektrums kann dadurch erklärt werden, daß bei Vergrößerung der Mundöffnung alle Oberwellen des Vokaltraktrohres weniger gedämpft werden (fig.2) und der normalerweise dreieckförmige Glottisimpuls durch Muskeltraining sich in Richtung des obertonreicheren Rechteckimpulses entwickelt.

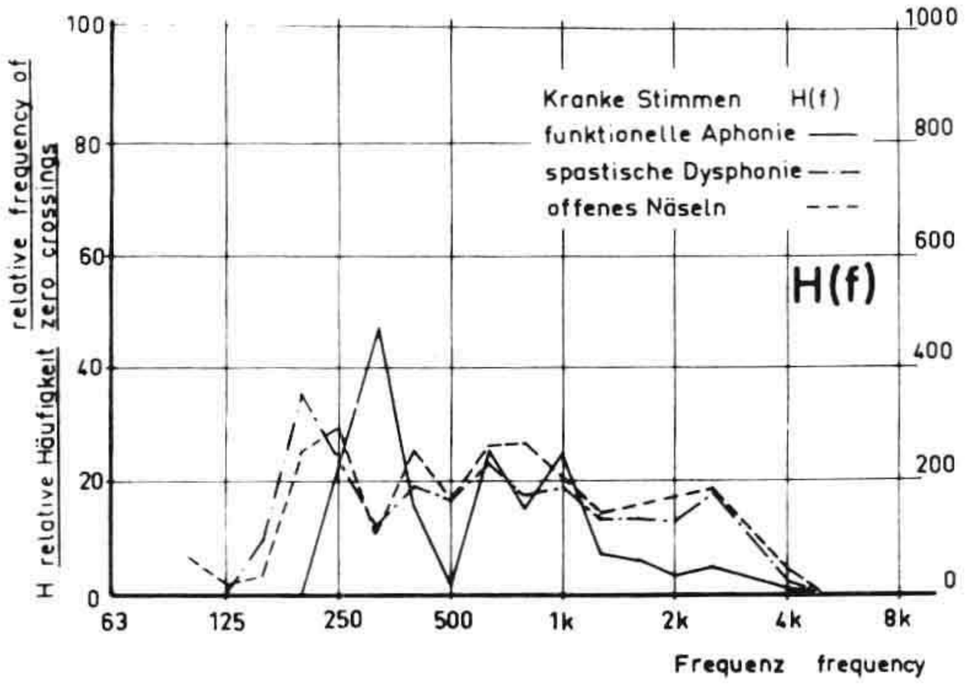


Fig.3 Verteilung der Nulldurchgangsrate für einen japanischen Sprecher ___japanisch --- deutsch gesprochen Beachte die Abweichung zwischen 1600 und 3500 Hz

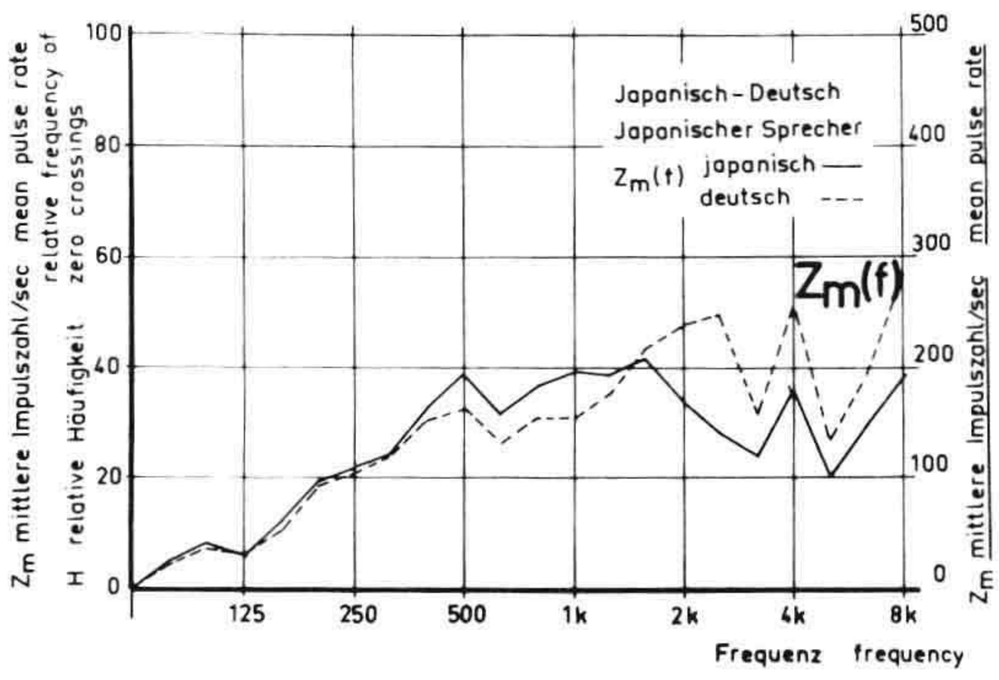


Fig.4 Häufigkeitsverteilung kranker Stimmen

Da im Häufigkeitsspektrum für die Mehrzahl der Vokale die F_3-Formanten sich in einem gemeinsamen Frequenzbereich überlagern, erklärt sich die hohe Zählrate. Daraus leitet sich der diagnostische Wert des F_3-Bereichs her. In einer weiteren Untersuchung sollte festgestellt werden, ob ein Sprecher, der in verschiedenen Sprachen spricht, ein unterschiedliches Verhalten zeigt. Für einen japanischen Sprecher stimmt der Kurvenverlauf für F_1 und F_2 bei deutschem und japanischem Text weitgehend überein (fig.3), lediglich im Bereich 1600 bis 3000Hz ergibt sich für den japanischen Text ein stärkerer Abfall, was den 3. Formanten aller Vokale betrifft, außerdem den 2. Formanten der Vokale i, ɪ, ɛ, æ . Dies stimmt damit überein, daß im Japanischen der neutrale und auch der nasalierte Vokal fehlen. Anscheinend wird die japanische Artikulation von einem höheren Muskeltonus angetrieben als die deutsche. Die Ausprägung von F_4 deutet auf höheren Muskeltonus auch im Glottisgenerator.

Häufigkeitsanalysen von Sprachstörungen und Stimmerkrankungen zeigen vielfältige und abnorme Unterscheidungsmerkmale. Fig.4 zeigt, wie aus den Kurven Klassifizierungsmerkmale für Stimmkrankheiten entnommen werden können; z.B. erlaubt der kontinuierliche Kurvenabfall ab 800 bis 1000 Hz für funktionelle Aphonie eine Gegenüberstellung zur spastischen Dysphonie und zum offenen Näseln bei Lähmung des Gaumensegels. Die Sprechbehinderung zwingt den Patienten zur Tonussteigerung der Kehlkopfmuskulatur im Sinne der Erzielung eines schnellen und vollständigen Glottisschlusses (fig.4).

Ein weiteres Individualmerkmal läßt sich aus dem Vergleich der Häufigkeitsverteilungen bei zwei etwa 10-15 dB voneinander verschiedenen Schwellwertpegeln gewinnen. Die bei höherem Schwellwertpegel erfaßten Werte sind im Mittelbereich "quasistationäre" Anteile der Sprachlaute. Das Verhältnis beider Verteilungen ist ein Maß für die Sprechaktivität der Versuchsperson und kann als "Artikulationsgrad" bezeichnet werden. Er ist im Bereich von 350 bis 4000 Hz bei geschulten Sprechern um etwa 1,5 - 2mal größer als bei normalen Sprechern.

Über die mit gleicher Meßanordnung ausgeführten Kurzzeit-Lautanalysen und die Verarbeitung der digital anfallenden Meßwerte im Computer zur Lautklassifizierung wird an anderer Stelle berichtet werden.

Die Deutsche Forschungsgemeinschaft hat diese Arbeit finanziell unterstützt. Die Messungen wurden von Ing. Roswitha Oellerich ausgeführt.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-2-9

LA SEGMENTATION TEMPORELLE : FORME DE PRELEVEMENT ET PERTURBATION SPECTRALE.

Sample.

LANDERCY Albert
Assistant à l'Institut de Phonétique, Université de Bruxelles.

WAJSKOP Max
Professeur à l'Institut de Phonétique, Université de Bruxelles.

Introduction

Afin de vérifier s'il est possible de décomposer en éléments catégoriels ce phénomène continu qu'est l'onde de pression rayonnée par l'appareil vocal de l'homme, il est utile de pouvoir prélever, au sein de la chaîne parlée, une série d'échantillons déterminés sans y introduire de perturbations. Dans ce but, nous avons construit un appareillage composé de deux segmentateurs électroniques et de leurs organes de commande (I). Cet ensemble nous permet d'extraire en n'importe quel endroit d'une séquence sonore, un segment de longueur prédéterminée. La fonction de prélèvement f(t) agit avec des fronts d'ouverture et de fermeture en forme d'arcs exponentiels à constante de temps réglable.

Mode de fonctionnement de l'appareillage (2).(fig. I) Le signal sonore g(t) dont on veut prélever un segment est enregistré sur la piste supérieure d'une bande magnétique montée en boucle et est ensuite dirigé vers la porte électronique par des circuits d'adaptation. Sur la piste inférieure sont enregistrés une impulsion zéro suivie d'un train d'impulsions d'horloge (2000 Hz).

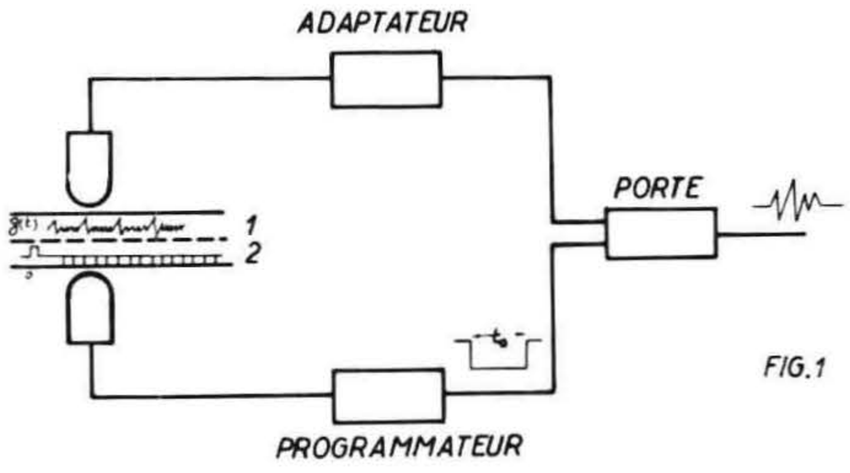


FIG.1

Dans ce train, nous prédéterminons deux impulsions n_I et n_2 ($n_I < n_2$). Le signal g(t) n'est transmis au travers de la porte que pendant le temps séparant ces impulsions. Nous dirigerons le signal de la piste inférieure vers un programmateur constitué d'un compteur décimal à précoups qui permet de prédéterminer n_I et n_2. A chaque passage de la bande, l'impulsion zéro remet le système dans son état initial.

Entre n_I et n_2, le programmateur délivre une impulsion dont la durée t_o correspond à la durée d'ouverture de la porte. Cette impulsion, après avoir été mise en forme par un circuit intégrateur qui réalise des fronts d'ouverture et de fermeture exponentiels, ouvre le segmentateur. A la sortie, nous recueillons donc le signal sonore g(t) modulé par la fonction de prélèvement f(t).

La fonction de prélèvement et son spectre. (3, 4, 5)

Lorsqu'on prélève dans le temps une partie quelconque d'un signal, la fonction de prélèvement perturbe le spectre défini de $-\infty$ à $+\infty$.

Comparons la perturbation obtenue par la fonction de prélèvement f(t) (I) avec celle créée par un prélèvement rectangulaire et essayons de définir une constante de temps d'ouverture et de fermeture optimale.

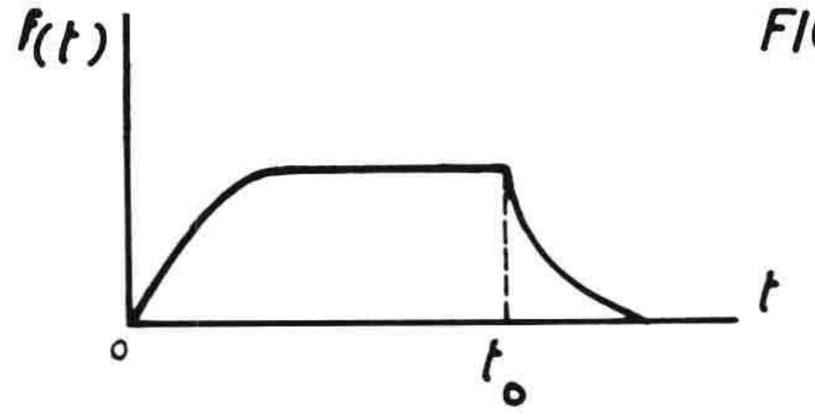


FIG.2

La fonction de prélèvement f(t) que nous utilisons peut être définie comme suit (fig. 2)

$$f(t)=\begin{cases} 0 & \text{si } t<0 \\ (I-e^{-\alpha t}) \quad (I) & 0\leq t\leq t_o \\ (I-e^{-\alpha t_o})(e^{-\alpha(t-t_o)}) & t_o<t<\infty. \end{cases}$$

Le spectre d'une fonction sinusoïdale $e^{j\omega_o t}$ modulée par f(t) s'écrira

$$|F(\omega_I)|^2 = \frac{\alpha}{\alpha t_o - I + \exp.(-\alpha t_o)} \quad \frac{\alpha^2 t_o^2}{(\omega_I^2 + \alpha^2)} \left| \frac{\sin(\omega_I t_o/2)}{\omega_I t_o/2} \right|^2 \qquad (2) \text{ avec } \omega_I = \omega - \omega_o$$

(fig. 3)

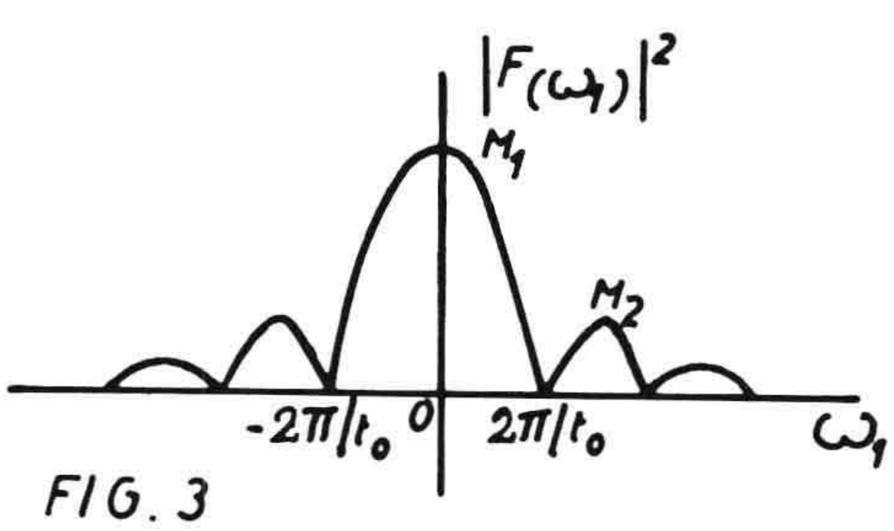


FIG. 3

$|F(\omega_I)|^2$ a la même allure que la fonction $(\frac{\sin x}{x})^2$ et présente donc des zéros pour $\omega_I = 2n\pi/t_o$ et des maxima $\omega_I = 0$ et $\omega_I = (2n+I)\pi/t_o$ (n=I, 2,...)
Si nous appelons respectivement M_I et M_2 les maxima en $\omega_I = 0$ et en $\omega_I = 3\pi/t_o$, nous pouvons calculer

$$M_I/M_2 = (3\pi/2)^2 \left[I + (3\pi/t_o)^2\right] \quad (3)$$

Si nous maintenons t_o constant et que nous faisons tendre α vers l'infini, nous obtenons une fonction de prélèvement rectangulaire. Dans ce cas M_I/M_2 est minimum et vaut $(3\pi/2)^2$. Nous constatons que M_I/M_2 sera toujours supérieur à $(3\pi/2)^2$ quelque soit α. Le fait d'employer une fonction de prélèvement telle que celle définie en (I) diminue donc l'importance relative du premier rebondissement spectral. Théoriquement M_I/M_2 sera maximum et infini lorsque $\alpha t_o \to \infty$. Mais une telle modulation ne représente plus aucune signification physique.

Afin de définir un $\alpha\ t_o$ optimum, nous avons adopté pour règle que 90% de l'énergie du signal soit transmise pendant le temps t_o.

Appelons E l'énergie transmise pendant la durée totale du signal et E_2 celle transmise pendant le temps $t > t_o$; nous devons résoudre l'équation $E_2 = 0{,}I\ E$. Si en première approximation nous posons $\exp.(-\alpha t_o) = 0$ nous obtenons $\alpha t_o = 6$. Dans ce cas, $M_I/M_2 = 77$.

Calculons la fréquence λ à laquelle l'équation (2) vaut $M_I/_2$. Cette fréquence permettra de calculer la largeur de bande à 3 db. L'équation (2) devient

$$x = \left|2k^2/(x^2 + k^2)\right|^{I/2} \sin x \quad (4) \quad \text{avec} \quad \begin{matrix} x = \lambda t_o/2 \\ k = \alpha t_o/2 \end{matrix}$$

équation qui peut se résoudre par approximations successives.

Un prélèvement exponentiel permet d'obtenir une largeur de bande inférieure à celle obtenue par prélèvement rectangulaire. Lorsque $\alpha t_o = 6$ on trouve, en résolvant (4) $\lambda t_o = 2{,}4$ tandis que si $\alpha\ t_o \to \infty$ on trouve $t_o = 2{,}8$.

Calculons le pourcentage d'énergie spectrale compris entre les deux premiers zéros du spectre (bande centrale). Il suffit de faire le rapport des deux intégrales $\int_{-\infty}^{+\infty} F^2(\omega_I)d\omega_I$ et $\int_{-2\pi/t_o}^{+2\pi/t_o} F^2(\omega_I)d\omega_I$.

Le pourcentage d'énergie spectrale compris dans la bande centrale est toujours plus élevé pour le prélèvement exponentiel que pour le prélèvement rectangulaire. Lorsque $\alpha t_o = 6$, nous trouvons 98% de l'énergie compris dans cette bande tandis que si $\alpha t_o \to \infty$, ce pourcentage est réduit à 90.

En conclusion, si nous comparons les spectres prélevés à l'aide d'une impulsion à fronts raides à ceux prélevés à l'aide d'une impulsion à fronts exponentiels, nous constatons que :

I. Les maxima et zéro sont situés aux mêmes fréquences dans les 2 cas.
2. M_I/M_2 est toujours supérieur pour le prélèvement exponentiel.
3. La largeur de la bande centrale est toujours inférieure pour ce même prélèvement.
4. La quantité d'énergie spectrale répartie dans la bande centrale y est toujours supérieure.

Diverses expériences de psychoacoustique en cours de réalisation tendent à prouver qu'à t_o identique, la perturbation spectrale n'est pas perçue lors du

prélèvement exponentiel alors qu'elle est nettement audible lors du prélèvement rectangulaire.

Références.

(I) Landercy, Sylin, Wajskop - La segmentation temporelle.
Rapport d'activité n° I, Institut de Phonétique - U.L.B. I966 - Bruxelles.
(2) Landercy, Sylin, Wajskop - Etude et réalisation d'un segmentateur électronique et son organe de commande. (sous presse) - La Revue d'Acoustique (France).
(3) Borg - Théorie des circuits impulsionels. C.N.E.T. - I953.
(4) Panter : Modulation, Noise and Spectral Analysis. McGraw Hill - I965.
(5) Pimonov - Vibrations en régime transitoire. Dunod. I962.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-2-10

Statistical properties of individual voices

Bordone-Sacerdote Cesarina
Sacerdote Gino
Istituto Elettrotecnico Nazionale - Torino (Italy)

1.- The voice of any single subject may be characterized from many points of view, according to different parameters depending on the sex, the age, the anathomical characteristics of the vocal apparatus, on the emotional conditions of the subject, on the language he speaks.

The voice of a subject may be considered under the aspect of its melody, its rythm, its timbre, its intensity: the scope of our investigation is to find an individual characteristic, in order to divide voices into particular cathegories.

As a first tentative we have investigated the power spectrum of the voice (a parameter already used by different authors, f.i. by Halle (1)) of subjects homogeneous for sex, age, language and condition.

In order to have any voice in form of a stationary phenomenon, we have used a "choral" method similar to the one employed by Tarnoczy (2): we have mixed the voice of the same subject by means of an 8 tracks magnetic recorder on an endless loop 10 seconds long. The subject reads in a plane way a text of fiction: the microphone output is successively recorded on eachone of the 8 tracks, whose outputs go to a mixer that can sum the instantaneous values of any number of tracks, from 1 to 8. The signal at the output of the mixer is as more uniform as the number of summed tracks increases: fig. 1 shows the signal of 1 tracks (a) and 2 (b), 4 (c) and 8 (d) tracks summed up.

2.- The statistical distribution of the relative level at the output of

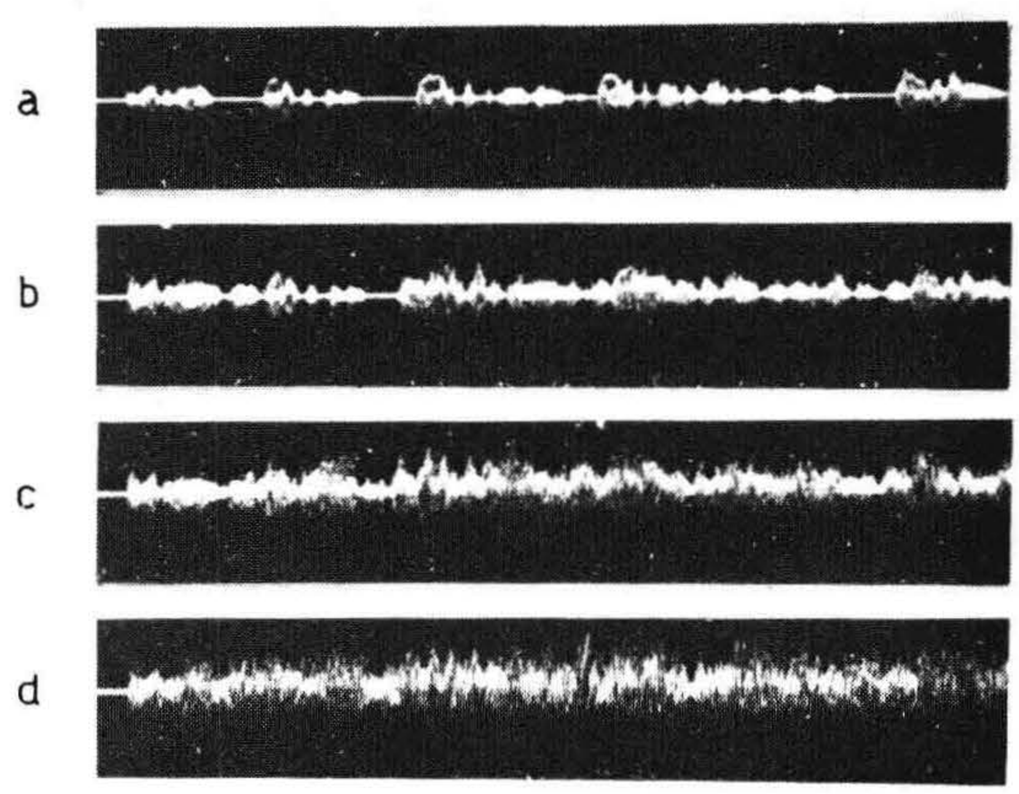


fig.1

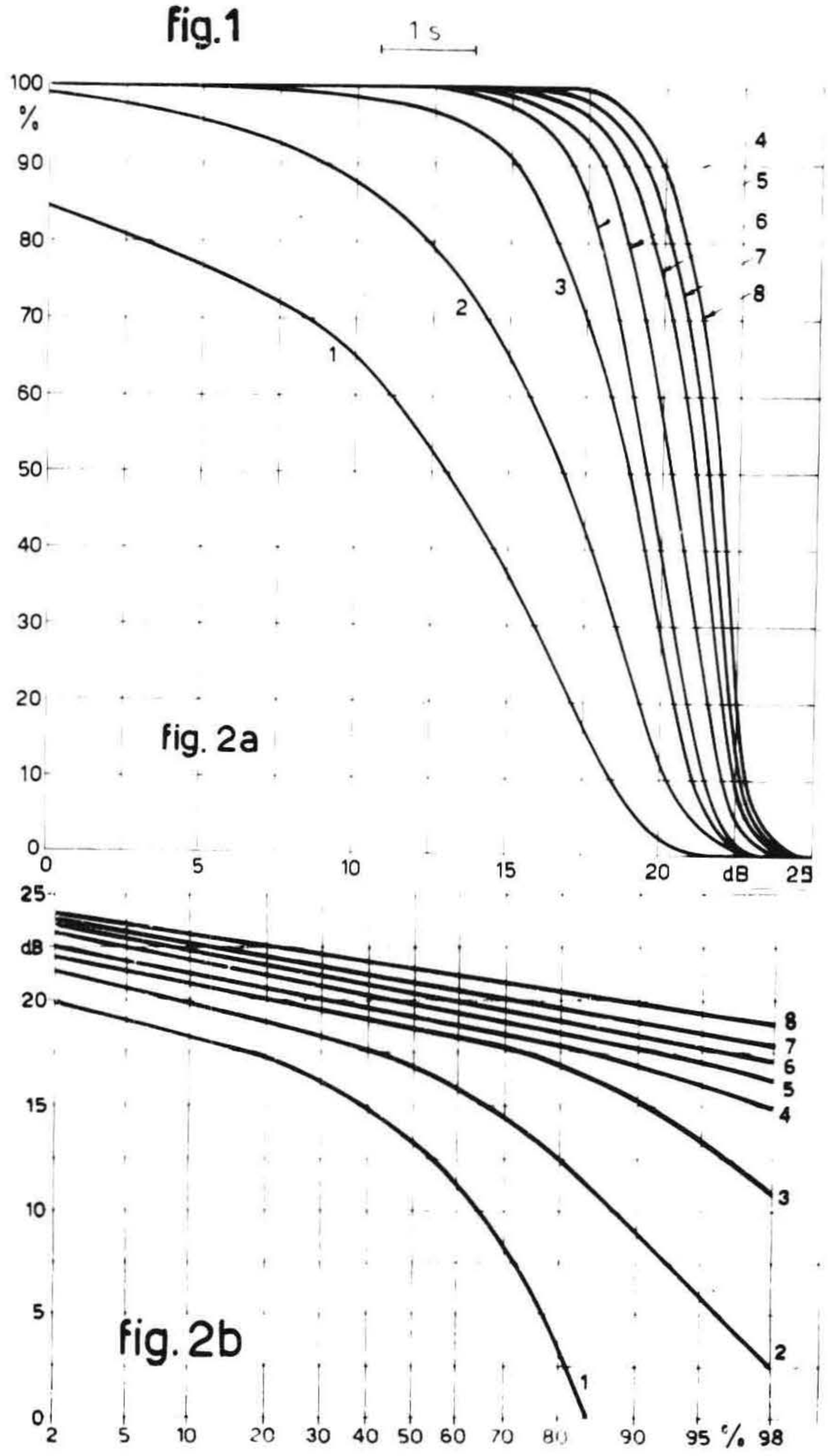


fig. 2a

fig. 2b

the mixer has been measured over 1000 samples counted every 0.1 second. The cumulative distributions measured for different numbers of tracks of the same loop (from 1 to 8) is reported in fig. 2 a: the same distribution is reported in fig. 2 b on a gaussian chart: on these coordinates if the diagram approximates a straight line the distribution is gaussian: in our case this is verified for a number of tracks greater than 5. The slope of these straight lines allows to deduce the r.m.s. dispersion , which decreases as the number of mixed tracks increases and tends to an asymptotic value that for different subjects varies from 1.3 to 1.8 dB (only exceptionally reaching 2.2 dB).

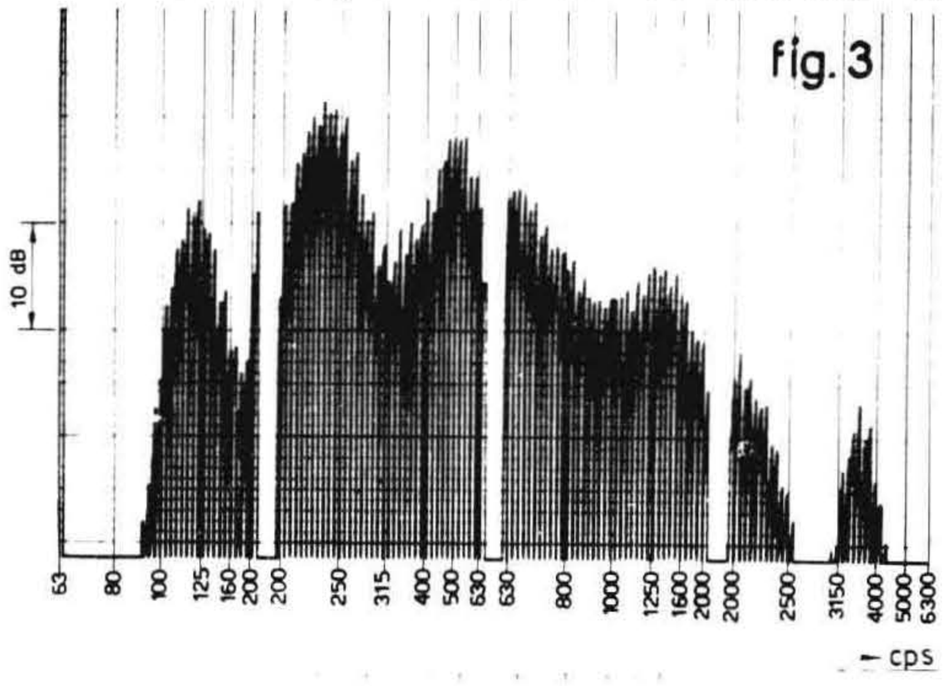
For a signal filtered by 1/3 octave band the distribution of the relative levels is also gaussian, but the asymptotic value of increases up to about 3 dB.

Within this approximation an 8 tracks signal may be considered as sufficiently stationary.

3.- The recorded power spectrum (P.S.) is obtained by an analy-

sis with a filter whose constant bandwidth is 1/3 octave. The nominal frequency of the filter changes continuously and automatically very slow ly. The filter output is squared by a simulating circuit in 5 steps of polarized current generators and integrated by an operational circuit. The reset is synchronous with the endless loop. Fig. 3 reports one of these P.S.. For the same subject the differences found for tests repeat ed many times on different days are contained in a narrow band around a curve that draws the envelope of the P.S. of the subject (fig. 4 a and b). Fig. 5 reports the P.S. of 9 voices in form of patterns.
A further scope will be to divide these patterns into cathegories.

fig. 3



The emotive condition of a sub ject (in connection f.i. with what he reads) affects greatly his P.S.: fig. 6 reports the patterns concern

fig. 4a

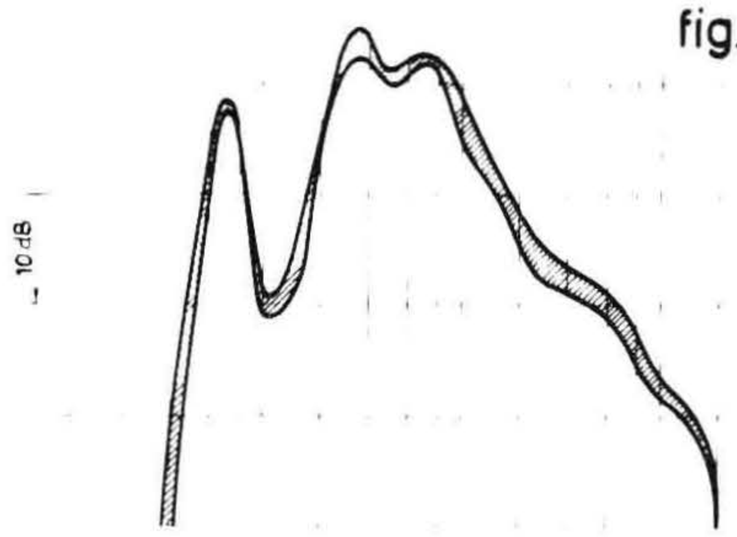
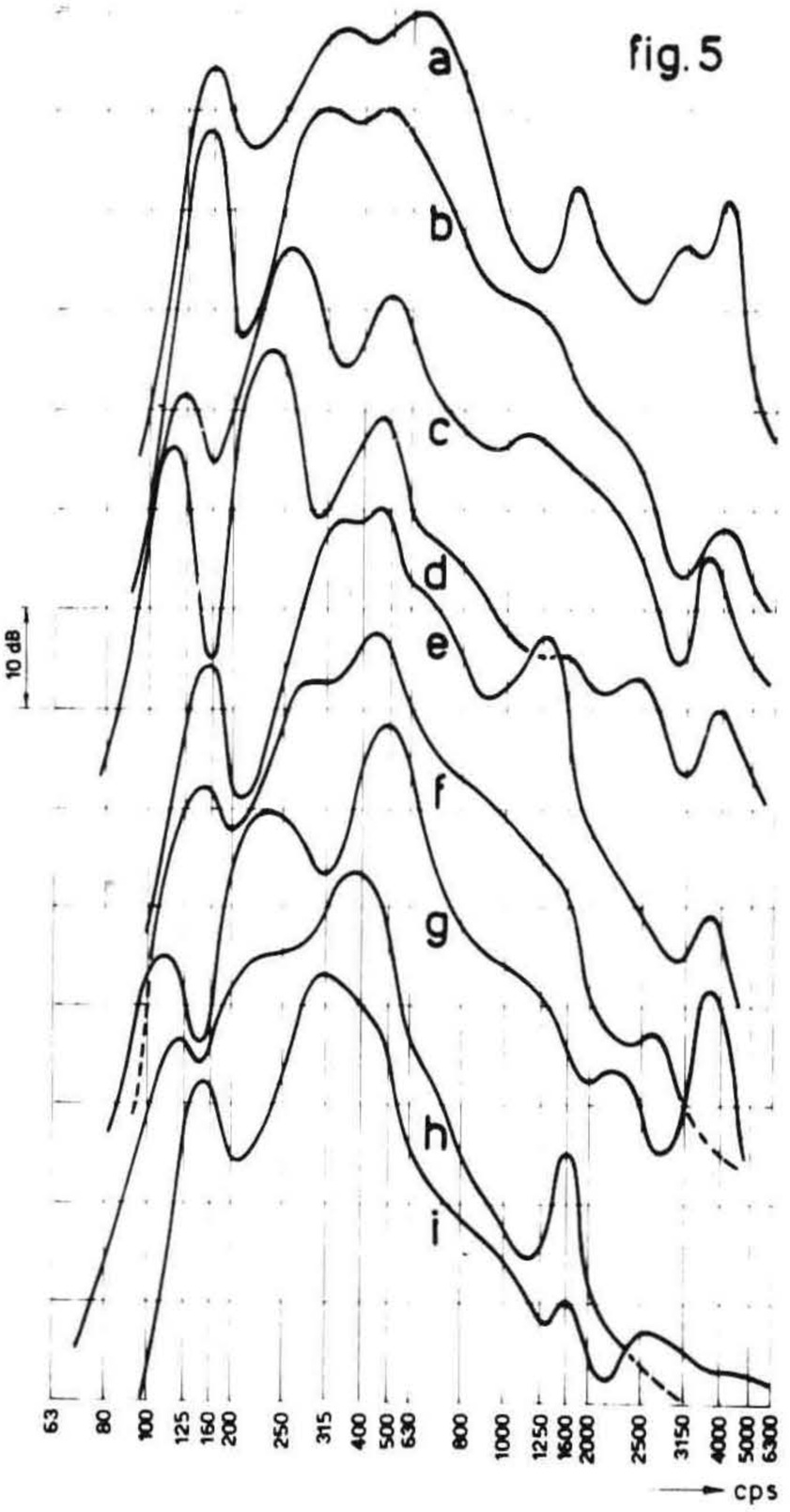
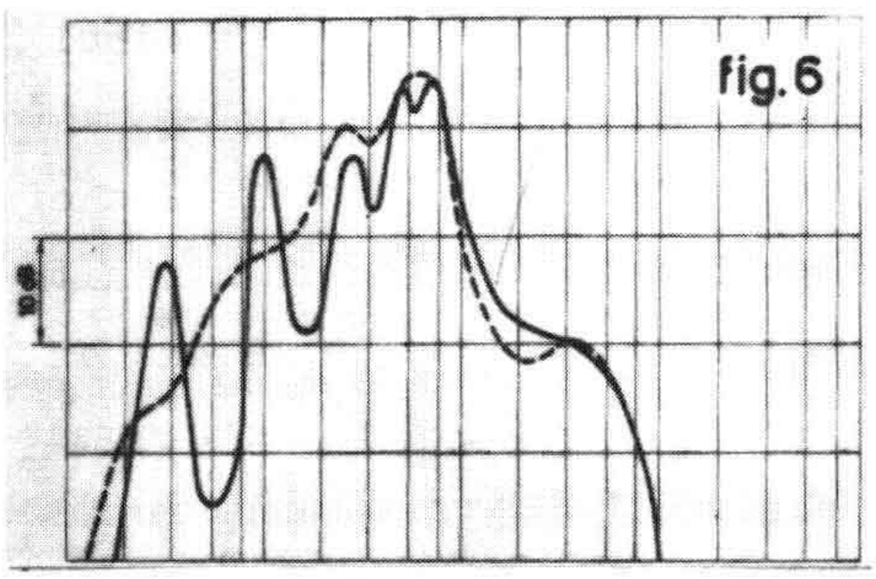


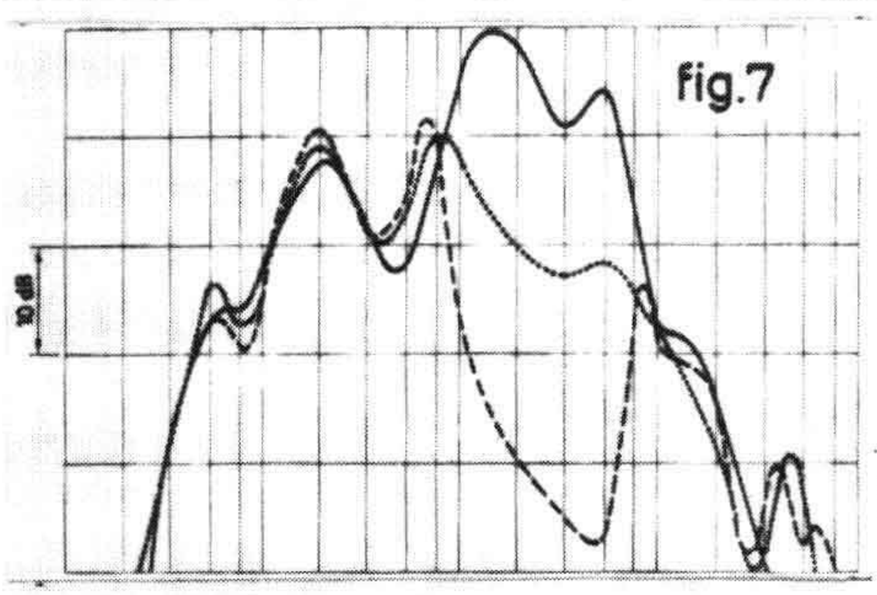
fig. 4b

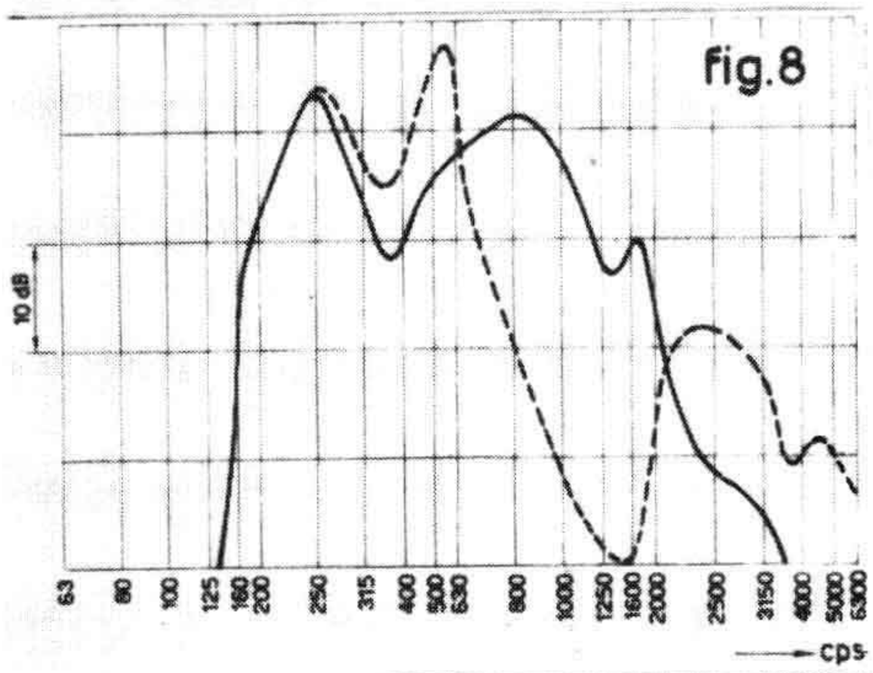


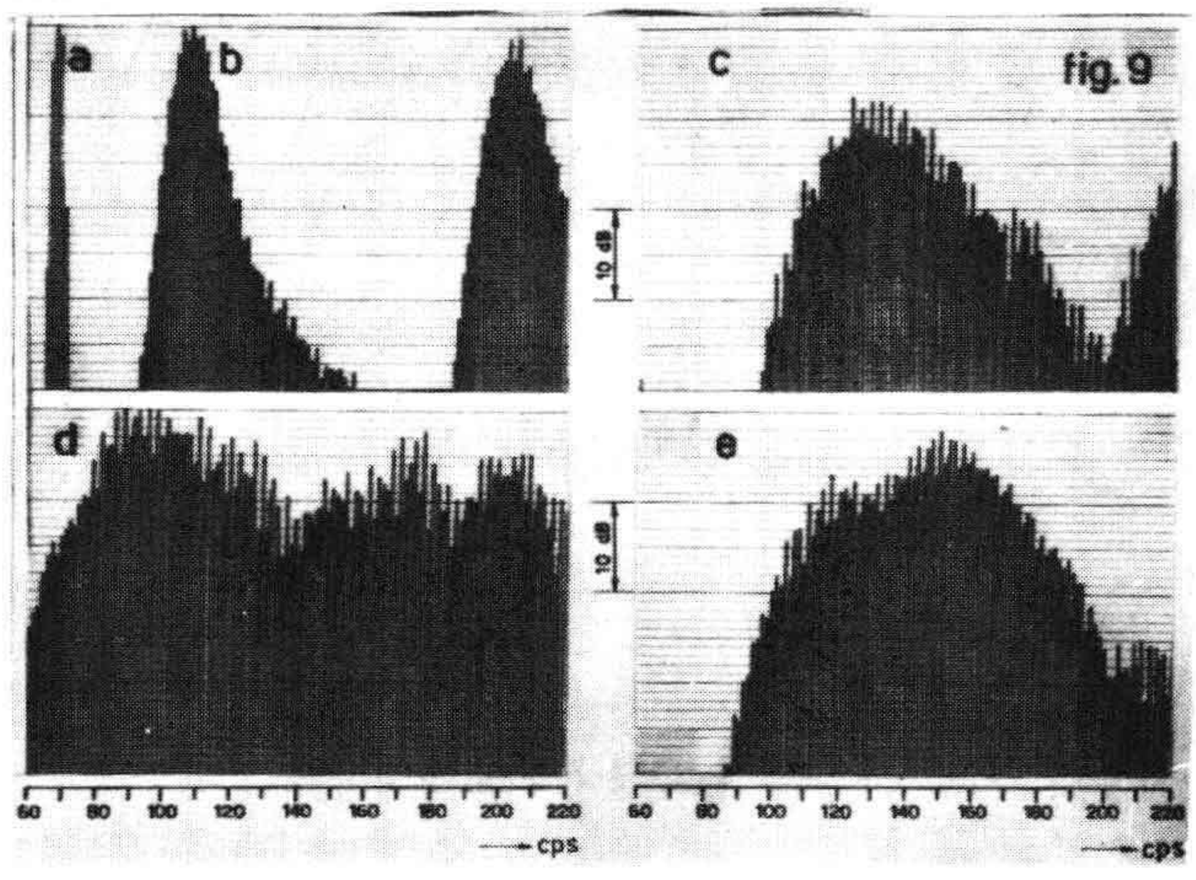
fig. 5











ing a very monotonous text (continuous line) and an usual text of fiction (slotted line), read by the same subject.

These patterns show that most components above the fundamental and the second harmonic correspond to the formants of a and e, which are the most frequent vowels in italian language. Using a text where only vowels a or e are used (f.i. Anna vara la barca, alza la randa...) we have obtained the patterns of fig. 7, in which the continuous line concerns a text with only a and the slotted line a text with only e. The dotted line is an usual pattern.

Fig. 8 reports the same data for loops in which every track concerns a different subject.

4.- To investigate the distribution of the fundamental of a voice, we have used a similar method, only the filter has a constant width of 3 cps. Fig. 9 reports some examples concerning one pure tone (a) and different voices (b, c, d and e).

(1) M.Halle - The sound pattern of Russian - Mouton Ed. 1959, p. 117.

(2) T.Tarnoczy - Détermination du spectre... - Acustica, VIII, p. 392.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

SEUILS DE DETECTION DE VOCOÏDES ISOLES EN FRANCAIS. †

WAJSKOP Max
Professeur à l'Institut de Phonétique
Université Libre de Bruxelles.

Introduction

D'une manière générale, les études sur le seuil de détection auditif ont été effectuées à l'aide de stimuli tels que les sons purs, les bruits blancs ou les trains d'impulsions. Les sons du langage ont été utilisés pour les recherches sur l'intelligibilité, soit en audiologie, soit en téléphonométrie. Parce qu'ils sont d'une manipulation difficile, ils n'ont pas été retenus comme stimuli de détection. Cependant, la voyelle, même isolée de son contexte, présente des traits propres à retenir l'attention. Elle est en effet un support virtuel d'information sémantique, elle s'insère dans un système linguistique d'oppositions distinctives, elle présente, sur le plan acoustique, une structure quasi-périodique liée à des résonances qui sont à la fois l'image de la fonction de transfert du conduit vocal et la base de cet attribut multidimentionnel qu'est le timbre sur le plan perceptif.

Le travail décrit ci-après est le second volet d'un programme consacré au traitement temporel du message verbal. Dans cette étude, nous avons voulu déterminer le seuil d'audition des sept voyelles fondamentales du français (I), isolées et présentées à durée constante.

Procédures expérimentales

Les voyelles ont été prononcées par un locuteur masculin, de manière soutenue, à intensité et à fréquence fondamentale constantes. Après contrôles auditifs et spectrographiques, des segments d'une durée de IOO msec. ont été découpés à l'aide d'une porte électronique (2,3). L'intensité de ces segments a ensuite été égalisée (entre 97,5 et IOO db). Les IO sujets ont été sélectionnés grâce à un audiogramme

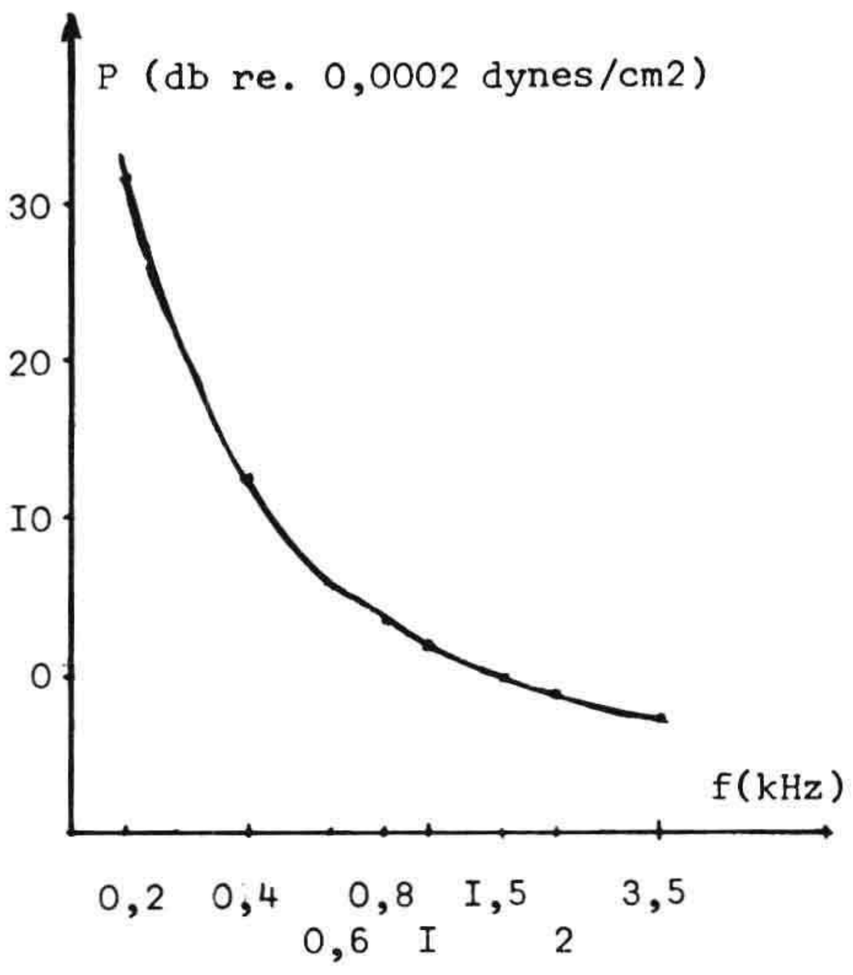


Fig. I

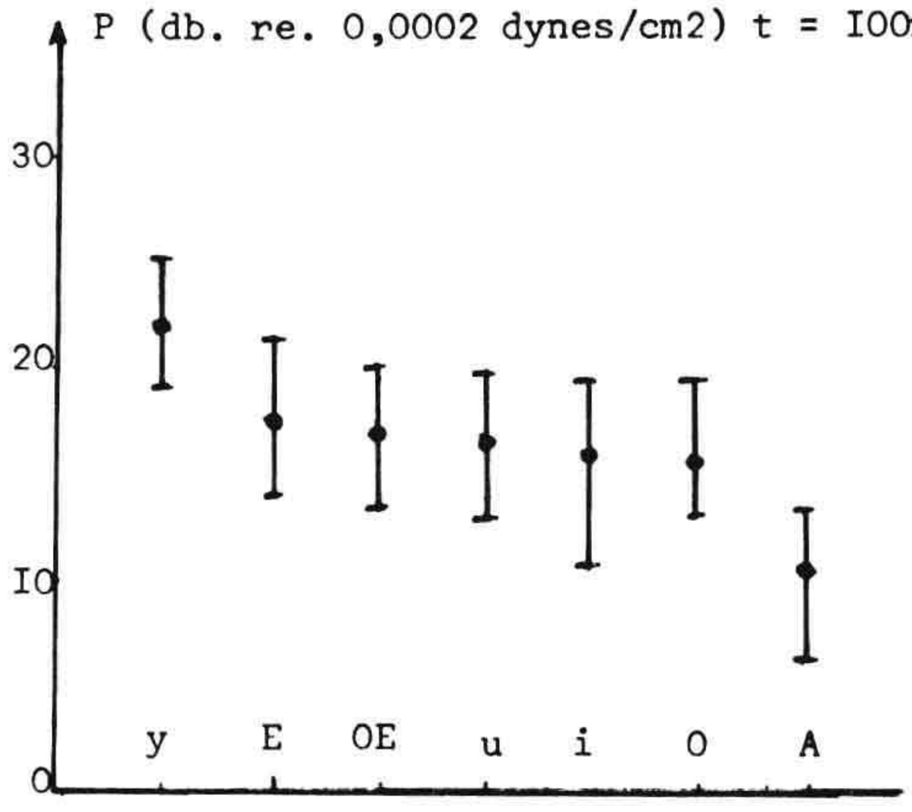


Fig. 2

biauriculaire sur des fréquences de 200 à 3500 Hz en utilisant la méthode des limites (I0 mesures par sujet). Cette mesure a été effectuée en chambre sourde. La fig.I illustre la moyenne des seuils.

La même procédure a été utilisée pour les voyelles, copiées sur bande magnétique à des intervalles de 2,5 sec. en séries de I5 stimuli. Dans chaque série, l'intensité des stimuli était identique à $\pm$ 0,5 db.

Résultats

Pour chaque sujet et chaque voyelle, on a calculé la moyenne des I0 mesures; on a ensuite extrait la moyenne des I0 sujets pour chaque voyelle (fig. 2).

On voit apparaître à l'examen visuel trois limites de détection concernant les voyelles [a] - [i, 0, u, 0E, E] - [y]. Ceci est confirmé par l'analyse de la variance (F calc. = 29,872 > $F^{.95}_{6;54}$ = 2,29). Enfin pour approfondir l'effet voyelles sur l'ordre des variations liminaires, on a utilisé une méthode de comparaison multiple S (4) qui confirme que les seuils des trois groupes de voyelles sont statistiquement différents à cause des valeurs de [a] et de[y] (5).

Cependant, les résultats individuels suggèrent l'hétérogénéité du groupe de sujets : F calc. = 3,00 > $F^{.95}_{9;45}$ = 2,05 (tableau I).

Seuils de détection de vocoïdes isolés en français.

Sujets	$\overline{M}_y$	$\overline{M}_E$	$\overline{M}_{OE}$	$\overline{M}_u$	$\overline{M}_i$	$\overline{M}_o$	$\overline{M}_a$
S_I	20,20	I9,30	I4,50	I5,60	I4,60	I4,00	9,00
S_2	2I,40	I8,80	I8,60	I6,80	I6,20	I5,90	I0,80
S_3	2I,00	I8,80	I6,60	I9,90	I3,50	I7,00	9,I0
S_4	22,00	I5,80	I7,00	I4,80	I8,40	I4,50	I0,40
S_5	I9,50	I4,I0	I5,70	I5,50	I6,30	I6,50	I3,30
S_6	20,70	I4,30	I9,70	I2,90	I6,40	I6,20	6,30
S_7	23,40	I6,70	I7,I0	I5,60	I0,60	I4,80	8,70
S_8	22,80	I4,30	I3,80	I7,30	I9,40	I3,20	I3,I0
S_9	25,30	I8,20	I5,00	I7,20	I4,50	I3,50	I0,70
S_{I0}	25,20	2I,80	20,00	I9,60	I9,30	I9,30	I2,50
$\overline{M}$ I00 mes.	22,I5	I7,2I	I6,80	I6,52	I5,92	I5,49	I0,39

Tableau I : Moyenne en db des seuils d'audition des 7 voyelles pour chacun des I0 sujets (I0 mesures par voyelle et par sujet).

Discussion

Toutes les conditions étant restées identiques : même fondamentale, même durée, même intensité à l'origine, même schéma de présentation, il résulte que la seule explication de cette différence au niveau de la détection réside I° dans la distribution de la puissance spectrale et 2° dans les capacités d'audition liminaire. Ceci confirme indirectement quelques-unes des observations énoncées par W.R.GARNER (6) pour des sons purs.Un contrôle quantitatif effectué après calibration du Sona-Graph, indique que la puissance spectrale relevée au seuil de détection est identique à $\pm$ 0,5 db aux résultats mesurés lors de l'expérience. D'autres vérifications sont en cours qui utilisent des méthodes psychophysiques différentes et

des stimuli plus longs afin d'éviter les effets de l'intégration temporelle (7,8)

Conclusion.

Les seuils de détection de 7 voyelles françaises isolées, égalisées en intensité et présentées à une durée constante de I00 msec. sont statistiquement différents. Cette différenciation provient de la répartition de l'énergie spectrale et est liée à la courbe de l'audiogramme tonal moyen.

Références

(I) B. MALMBERG - La structure phonétique de quelques langues romanes.
ORBIS, XI, I, pp. I3I-I78 (I962)

(2) LANDERCY A., G. SYLIN, M. WAJSKOP - La segmentation temporelle I.
Rapport d'Activités I, Institut de Phonétique, pp. 99;I35 (I966)

(3) A. LANDERCY, G. SYLIN, M. WAJSKOP - Etude et réalisation d'un segmentateur électronique et de son organe de commande.
La Revue d'Acoustique, G.A.L.F. Paris (à paraître).

(4) SCHEFE - The Analysis of Variance - New-York, Wiley, I959.

(5) La valeur critique de S$\hat{\Psi}$ = 2,3

(6) W.R. GARNER - The effect of frequency spectrum on temporal integration of energy in the ear.
J. Acoustic. Soc. Am., I9, pp. 808-8I5 (I947)

(7) W.R. GARNER, G.A. MILLER - The masked threshold of pure tones as a function of duration.
J. exp. Psychol., 37, pp. 293-303 (I947)

(8) W.A. MUNSON - The growth of auditory sensation.
J. Acoustic. Soc. Am. I9, pp. 584-59I (I947)

† Recherche subsidiée partiellement par le Fonds de la Recherche Fondamentale Collective, contrat n° 6I2.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-3-2

Harmonic Components and vowel Perception

Nobuo Hirato, Osamu Kakusho, Kanemasa Kato
Institute of Scientific and Industrial Research,
Osaka University, Osaka, Japan

Takashi Kobayashi
College of General Education, Osaka University,
Osaka, Japan

Introduction

Acoustic parameters mostly used at present in the field of speech study are very useful, indeed; but it has not necessarily been made very clear how they are related with auditory response. In studying speech as a kind of pattern recognition, it is probably very appropriate to look for parameters to which the auditory system may attend, i.e. parameters which play the most important role in the auditory system's response to the speech sound. There have been many excellent psycho-acoustic researches into the loudness, masking, or pitch of the sound having a comparatively simple spectrum, but not, it seems, as to the speech sound having a complicated spectrum. Considering the above two points, that is, in order to examine the auditory system's response to complicated sound wave such as vowels, we have taken up harmonic components of a vowel spectrum, and at first examined the DL (Difference limen) of each harmonic component. Then we examined the relations between the number of the largest harmonics and (1) phonemic and vocal qualities, (2) identification rate, and (3) pitch.

[1] DL of the amplitude of each harmonic component of a vowel

We chose the DL of the amplitude as a criterion to find which of harmonic components Hi (representing also the amplitude; i=1,2,...,32; fundamental frequency f_0=125Hz) of a vowel spectrum are attended to by the auditory system. Japanese five vowels were treated.

Experimental Method

(1) Standard Sound (cf. Fig. 1, which shows the case of /a/)

(i) $T(f) \times f^{-1}$ charactristic, shown as a thick solid line in Fig. 1, was formed by the terminal analog synthesizer. (ii) Each output Hi of the harmonic synthesizer is adjusted to the dot on the above-said solid

line. The spectrum of the standard sound thus synthesized is denoted by $P(f)$.

(2) Comparision Sound This is synthesized by changing a particular H_i of the standard sound, each of which difference is $\pm\Delta H_i$.

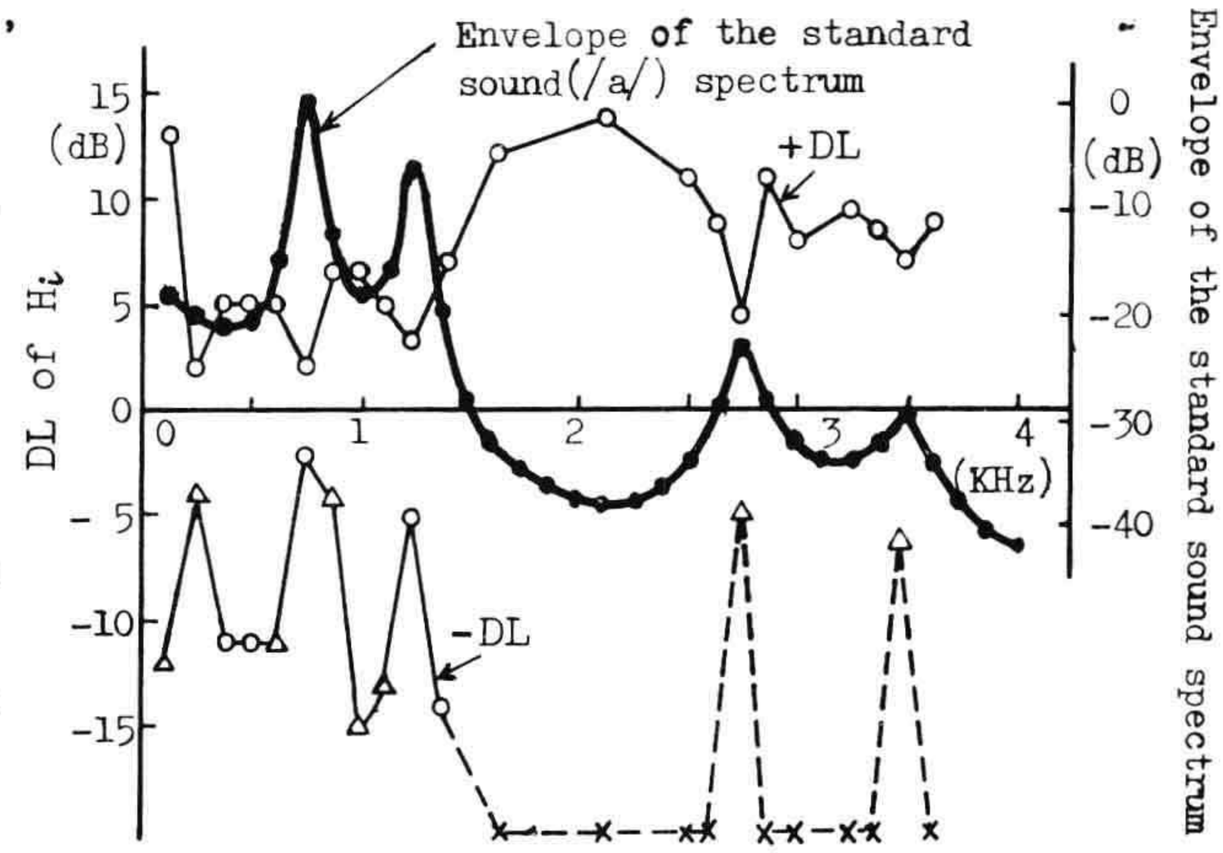
(3) Auditory tests are made by means of a constant method. N: the number of subjects; n: the number of a subject's trials per a test.

Experimental Results

In Fig. 1 is shown an example of +DL and -DL which are the results obtained by increasing and decreasing respectively each H_i. This reveals us the following: (a) +DL is small (2-5dB) as H_i is nearer to the formant frequencies (F_1, F_2, F_3) of the standard sound and the line connecting +DL's is just like the reverse of the frequency spectrum envelope of the standard sound ($E_{nv}[P(f)]$).

(b) As for -DL, it was immeasurable for most H_i. All the H_i whose -DL could be obtained from all the subjects are the ones that are near the formant frequencies and that have the maximum amplitude.

(c) Each DL cannot be regarded as resulting from a simple quantity like the difference of the total power of the sound.



● : amplitude and position of harmonic conponent H_i.
o : mean value of DL's of all subjects.
Δ : DL was immeasurable for one of the two subjects.
× : DL was immeasurable for both the two subjects.
N=2, n=20. SPL=65dB.

Fig. 1 DL of amplitude of each harmonic component H_i of Japanese vowel /a/.

[2]. Relations of the number and positions of harmonic components to (1) phonemic and vocal qualities, (2) identification rate, and (3) pitch

From [1], we can draw an inference that the H_i which has a smaller DL, i.e. the one which is nearer to the formant frequency and which has a larger amplitude, is more important. This leads us to make a series of synthesized sounds P_j ($j=1,2,\ldots,32$) for each vowel, where j means the number of harmonic components included in each synthetic sound and, as for the harmonic components included, those which compose the frequency spectrum of the standard vowel are added one by one in the order of decreasing amplitude. The frequency spectrum of P_j is denoted by $P_j(f)$.

The object of this experiment is to see what j will be when P_j is identified with P_{32} in terms of the following three: (1) phonemic and vocal qualities, (2) identification rate, and (3) pitch. (P_{32} is the standard sound, whose envelope of frequency spectrum is $T(f) \times f^{-1}$.)

Experimental Method

(1) Phonemic and vocal qualities

P_j and the standard sound P_{32} were made into AB pairs, which was done separately for each of the Japanese five vowels. By means of a constant method, we obtained the response rate R at which A and B were identified (i.e., the subject could not find any difference between A and B).

(2) Identification rate

P_j were made from the five vowels, and were arranged at random. Subjects were forced to identify each P_j with any one of the five vowels. The identification rate R_i is the rate at which each P_j was identified with the corresponding vowel.

(3) Pitch (cf. Fig. 2)

Subjects adjusted the fundamental frequency f_0 of $P(f)$ (synthesized by the matching pattern generator) to match that of $P_j(f)$ (synthesized by the set pattern generator).

This experiment was done for each vowel.

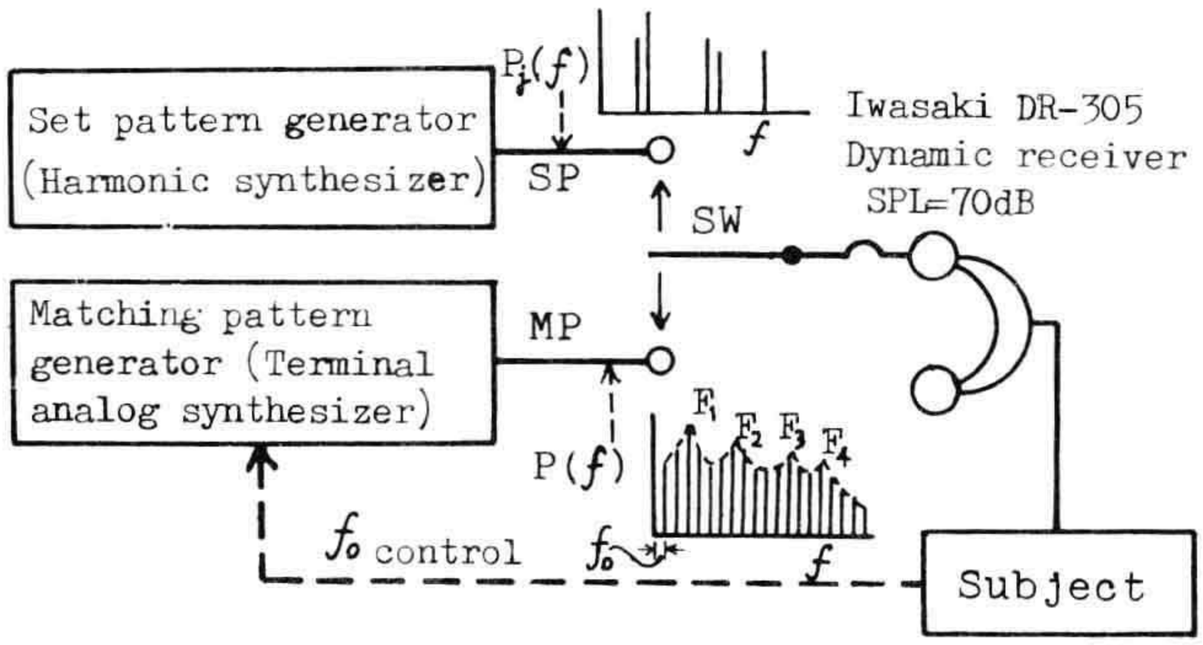


Fig. 2 Block diagram of pitch matching experiment.

Experimental Results

(1) The relations between j and R are shown in Fig. 3.

(2) In Fig. 4 are the relations between the identification rate R_i and the number of harmonic components j, and underneath are

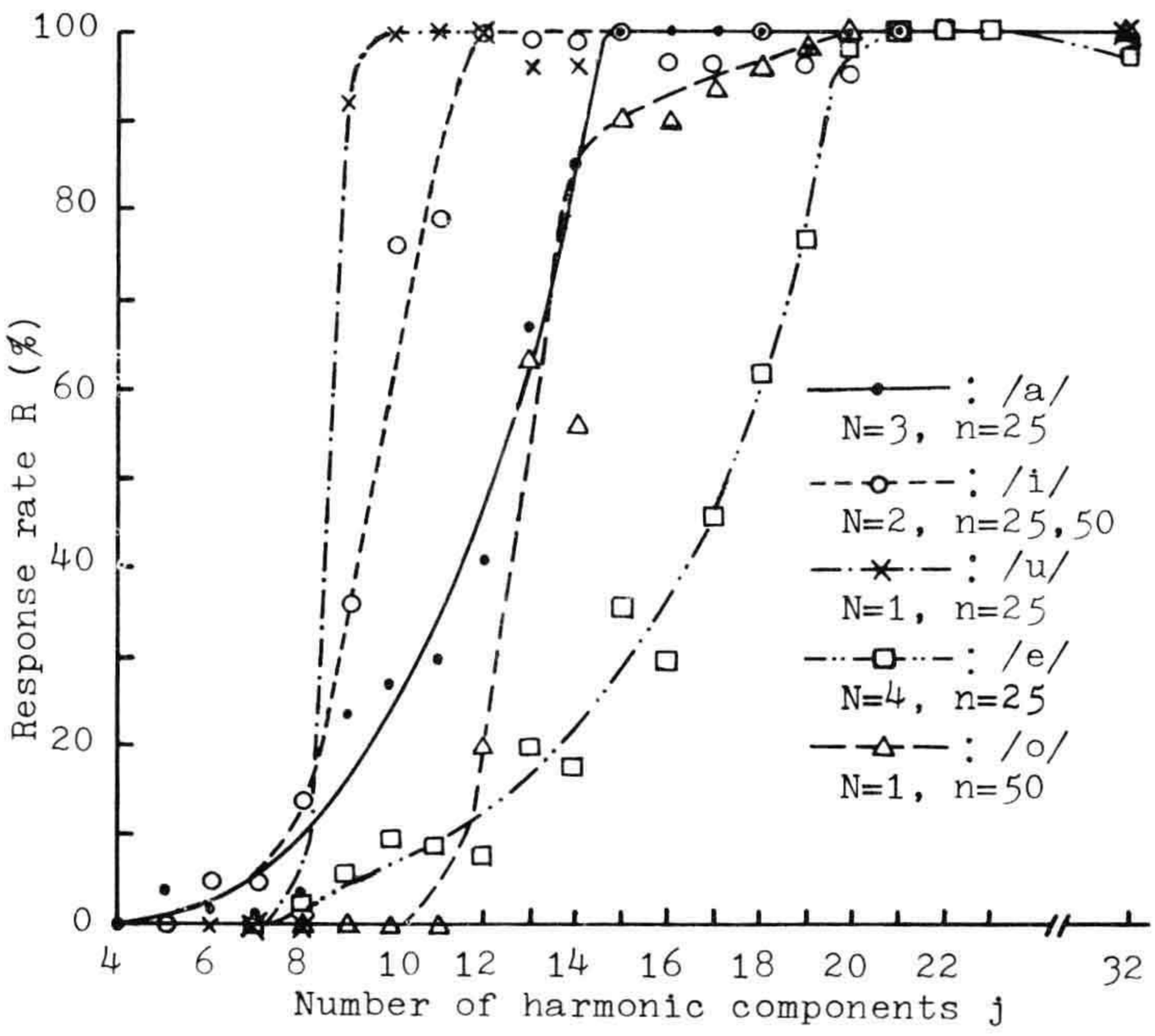


Fig. 3 Relations between the number of harmonic components j and the response rate R.

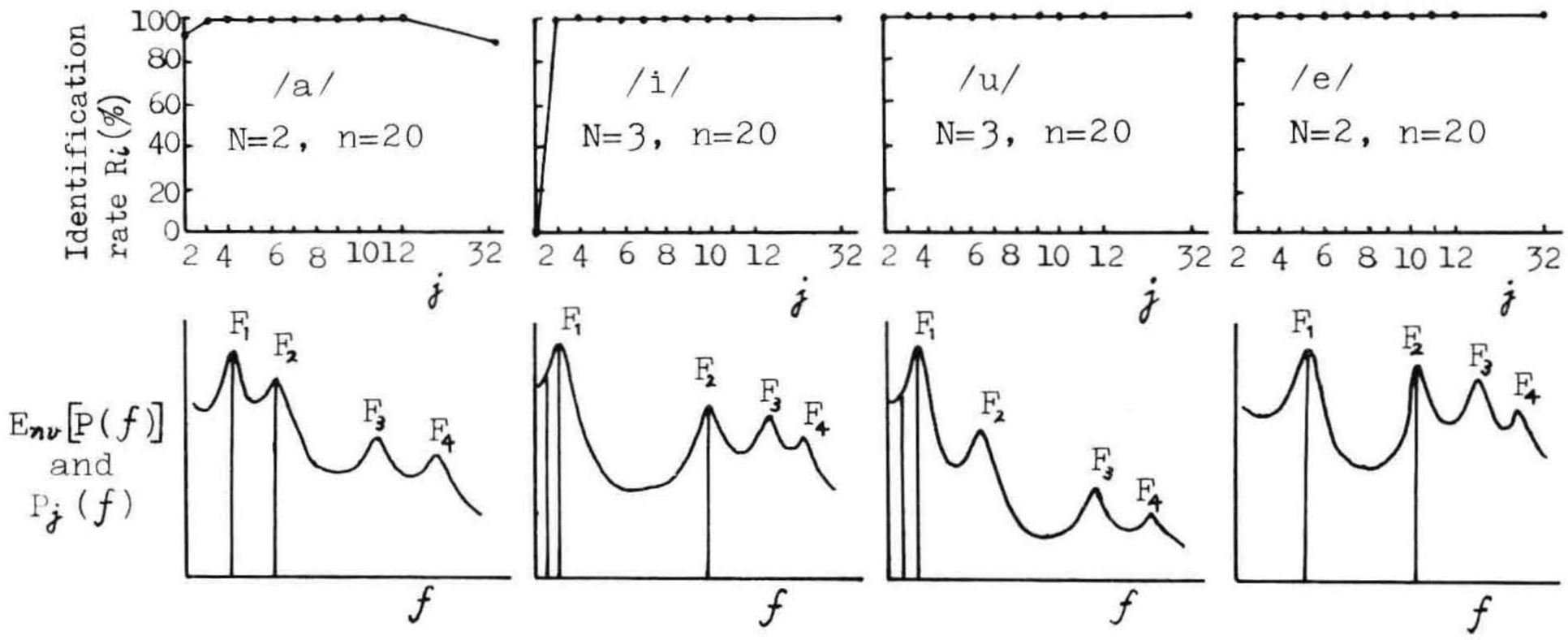


Fig. 4 Relations between identification rate R_i and the number of harmonic components j (above); and $P_j(f)$ corresponding to $R_i \geqq 90\%$ when $E_{n\nu}[P(f)]$ has the F_1 - F_2 combination of 100% identification (below).

shown the $P_j(f)$'s where $R_i \geqq 90\%$.

(3) Fig. 5 shows the relation between j and f_0.

[3] Conclusion

(a) So as not to deteriorate the phonemic and vocal qualities of our vowels, 9 (in the case of /u/) to 20 (in the case of /e/) harmonic components, whose amplitudes are larger than the others', are sufficient, and the number is approximately linear to $\sqrt[3]{F_1 F_2 F_3}$.

(b) Only two to three harmonics, which are near the formant frequency, are enough to get about a hundred per cent identification rate, if we choose for P_{32} the F_1 - F_2 combination corresponding to 100% identification. About twice as many harmonics are sufficient for a man to get correct information with regard to a fundamental frequency, i.e. to match f_0 to 125Hz in this case.

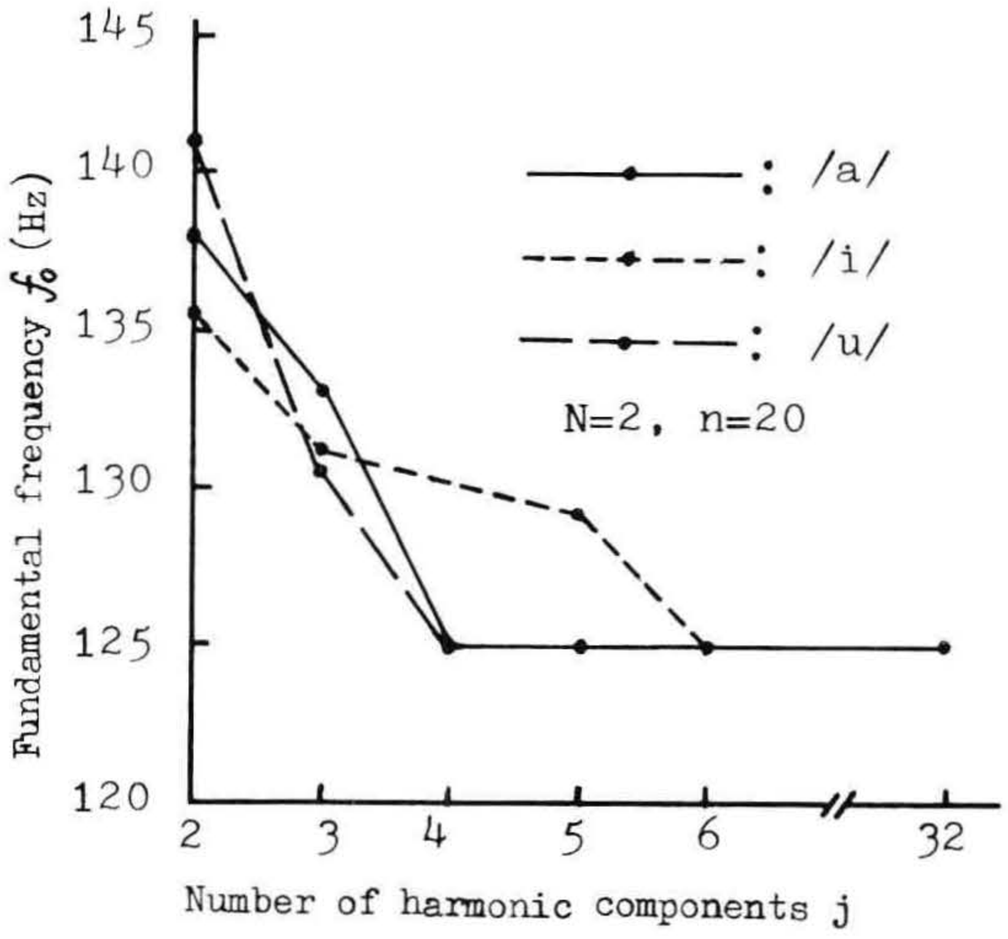


Fig. 5 Relations between the number of harmonic components j and fundamental frequency f_0.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-3-3

On Auditory Space Model of Vowel Perception

Hideki Kasuya
Research Assistant, Research Institute of Electrical Communication, Tohoku University
Hisayoshi Suzuki
Assistant Professor, Research Institute of Electrical Communication, Tohoku University
Ken'iti Kido
Professor, Research Institute of Electrical Communication, Tohoku University

Introduction

Much research has been carried out to find a rule for speech recognition acoustically, and it has cleared that a fairly high rate of recognition can be obtained with a stationary vowel, if the formant frequencies of each vowel sound is known. That is, with the knowledge of the sex and the age of the speaker, the stationary vowel can be well recognized by use of the 1st and 2nd formant frequencies only. It is also clear[1] that the vowel can be well recognized regardless with the sex or the age of the speakers, if the 3rd formant frequency or the voice fundamental frequency is also considered along with the 1st and 2nd formant frequencies.

However, it can hardly be supposed that human beings discriminate the vowels by the use of three dimensional space consisting of the first three formant frequencies. So, in this paper we attempt experimentally to make clear the manner of vowel perception of human beings. We shall call the space which represents the vowel quality "the auditory space" for the present. It is assumed that each vowel has its own distinctive position in the auditory space. To determine experimentally the structure of auditory space and to make a scale of the phonetic quality of the five Japanese vowels, the multidimensional scaling method is used. The relations between the auditory space and the formant frequencies are also discussed.

Experiments

In a scaling of the space Torgerson's complete triad method[2] was used, because of the clearness of criterion for judgement and the

stability of the experiments. The stimulus sounds, of which the formants and pitch frequencies are illustrated in Table 1, are synthetic vowels so that the acoustic parameters of the stimuli can be changed at will. Listeners are four girl students who have normal powers of hearing .

Results and Discussion

In general, if the distances between all pairs of n stimuli are given, n-1 dimensional space is enough to assign each stimuli without contradiction. As a result of experiment, the auditory space for assigning the Japanese vowels can be made with three dimensions. The configuration of vowels in the auditory space is illustrated in Fig. 1 with each of the listeners. The projections of the stimulus points on the axis were calulated using the theorem provided by Young and Housholder. The axis are named a_1, a_2 and a_3 according to their value of variance. The origin is moved without rotation of the axis so that /e/ is on the a_3 axis and $a_3 \geqslant 0$ for all vowels. The coordinate is rotated around a_3 axis so that /a/ is on the (a_1, a_3) plane.

In such a system, a_1 and a_2 correspond approximately to the distinctive features acute/grave and light/heavy[3] respectively. For a_3 it is difficult to see the significant feature common to all of the listeners, since it varies greatly with each listener. It is of much interest to know what of physical quantity is related to each axis.

The following assumptions are made: (1) a linear combination of logarithmic values of the frequencies of the first three formants is used in auditory perception, (2) relative values of formant frequencies are also used. As shown in Fig. 2, a_1 and a_2 can be approximated as follows from the average judgement of the 4 listeners.

$$a_1 = \log \frac{F_1^{4} F_3^{4.4}}{F_2^{8.4}} + 2.1 \quad (1) \qquad a_2 = \log \frac{F_3^{14.6}}{F_1^{9.1} F_2^{5.5}} - 6.1 \quad (2)$$

That is, the values of a_1 and a_2 are invariant for the frequencies of the three formants each ratio of which is kept constant. In spite of the changes of the formant frequencies with the sex and the age of the speakers the three formant frequencies satisfy approximately the condition described above. Therefore, Eqs. (1) and (2) can be regarded as the operators for normalization of the sex and the age factor. The F_1 versus F_2 representation of the formant frequencies of 8 years old children and male adults is illustrated in Fig. 3-a and the a_1 versus a_2 representation computed by equations from the same samples is also illustrated in Fig. 3-b. It can be seen that the acoustic parameters

F_1, F_2 and F_3 are markedly normalized on the (a_1, a_2) plane.

Acknowledgement

The authors wish to express their gratitude to Prof. S. Hiki and Prof. A. Watanabe, for their advice and encouragement, and to Y. Tanabe, for his valuable assistance in the experiments.

Reference

(1) H. Suzuki et al, Preprints of 1967 Conference on Speech Communication and Processing, B-5, Nov. (1967)

(2) W.S. Torgerson, Theory and Method of Scaling, John Wiley & Sons, Inc., London (1963)

(3) S. Hattori et al, Bull. of Kobayashi Inst. of Physical Res., 7, Japan, 69-79, (1957)

(4) G. Hanson, Ericsson Technics,No. 1, (1967)

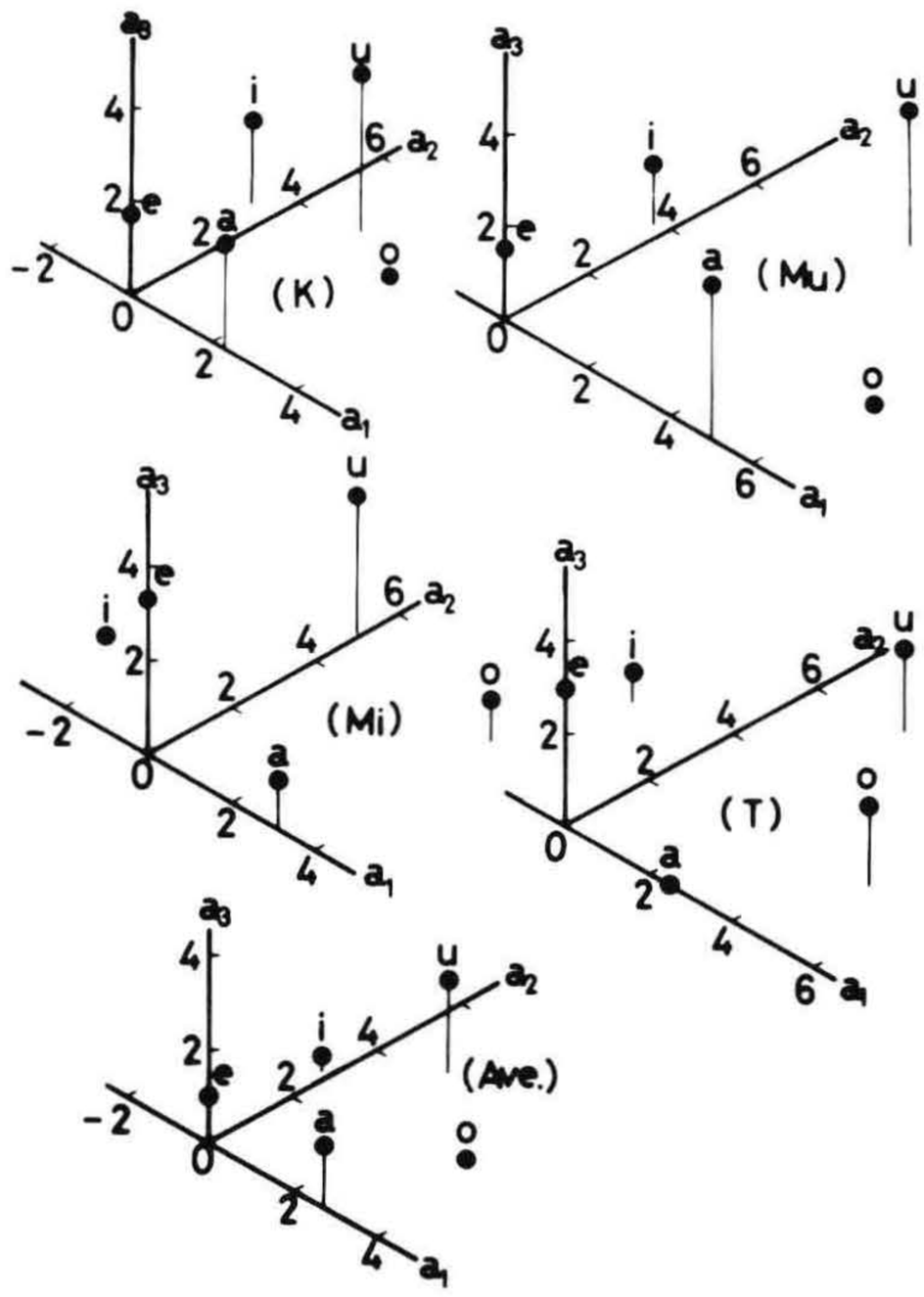


Fig.1 Configuration of the five vowels in the auditory space of each of the four listeners(K,Mu,Mi and T) and of the average of them(Ave.).

Tab.1 Formant and pitch frequencies of the test vowels

Vowel	F_0	F_1	F_2	F_3
/i/	104	300	2100	2400
/e/	104	500	1950	2400
/a/	104	700	1150	2400
/o/	104	500	740	2400
/u/	104	300	1150	2400

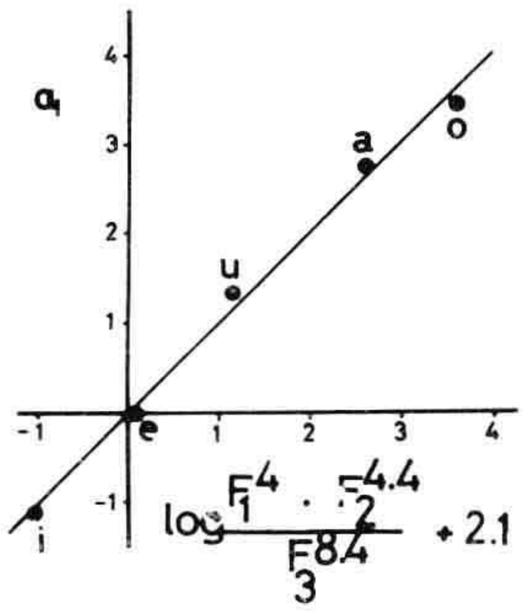


Fig.2-a Relation between frequencies of the first three formants and a_1

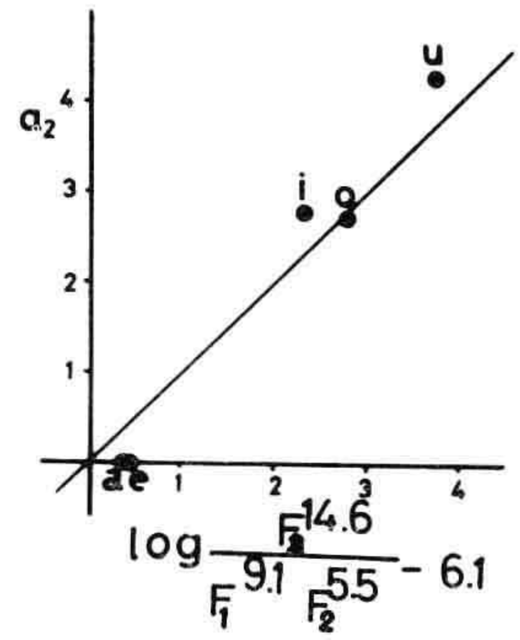


Fig.2-b Relation between frequencies of the first three formants and a_2

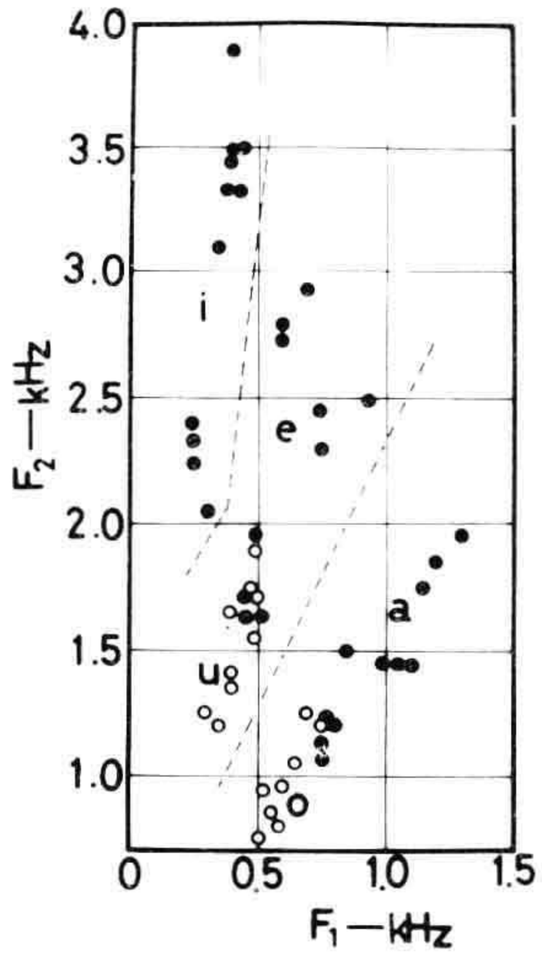


Fig.3-a First formant versus second formant representation of five Japanese vowels spoken by 8 years old children and male adults.

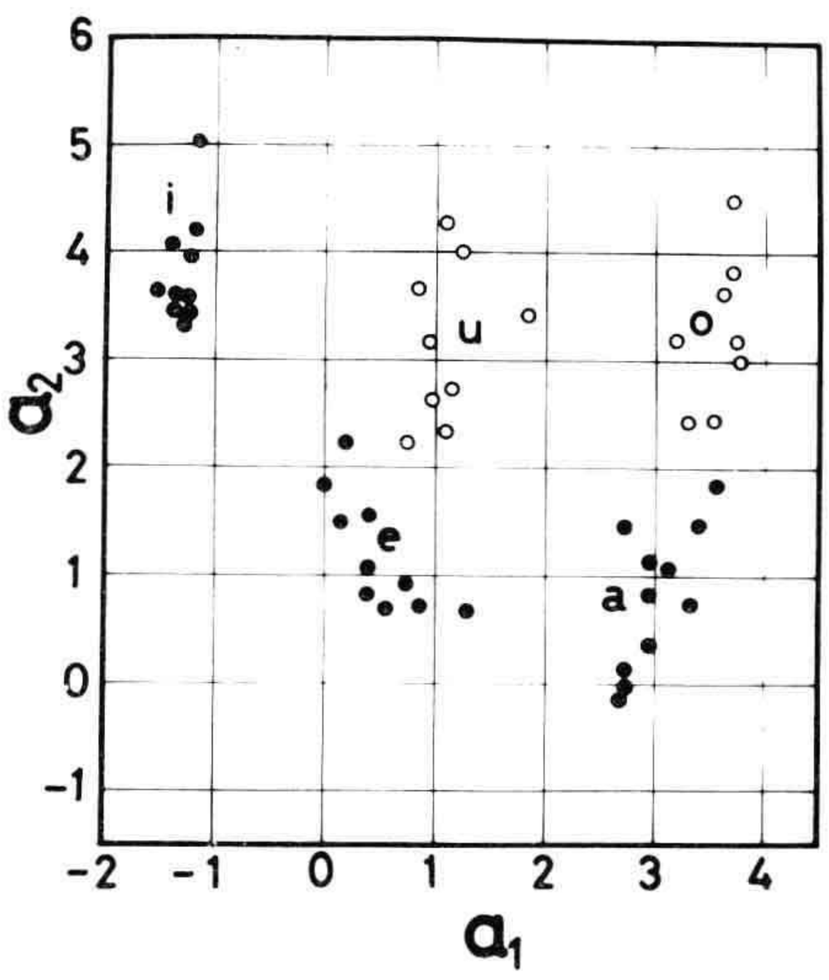


Fig.3-b a_1 versus a_2 representation of five Japanese vowels caluculated from the same samples with Fig.3-a.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-3-4

Dynamic Model of Vowel Perception

Shizuo Hiki, Hiroshi Sato, Takeo Igarashi and Juro Oizumi
Research Institute of Electrical Communication,
Tohoku University, Sendai, Japan

Introduction

In spite of the fact that the formant frequencies of the vowels in the connected speech are, in most cases, deviated considerably from those in the steady part of the monosyllabic vowels, the kind of vowels intended to speak by the talker can be identified correctly by the listener. One of the important reasons why the phenomena occurs is that the phonetic quality of the vowels is perceived relatively in the auditory system of the listener.

The process of the relative perception is considered to be composed of two steps. The first is the static perception according to the degree of the resemblance in the phonetic quality between the vowel stimulus and each one of five Japanese vowels, and the second is the dynamic perception according to the degree of difference in the phonetic quality between the vowel stimulus and the adjacent vowels in the connected speech.

In this paper, it is intended to formulate the model of the relative perception of the vowels in the connected speech through the results of the perceptual tests with the synthetic vowels.

Measurement

The test vowels were synthesized with the terminal analogue type speech synthesizing circuit and presented to the listeners binaurally through the dynamic receivers (DR 305) in the noiseless room. The subjects were five college students (female) and they were asked to assign each synthetic vowel to one of five Japanese vowels.

Static Model

The results of the perceptual tests with the synthetic vowels of various sets of the first and the second formant frequencies show that the probability of the identification of the synthetic vowel stimuli as Japanese /i/, /e/, /a/, /o/, or /u/ distributes widely around the standard vowels (refer Figure 1).

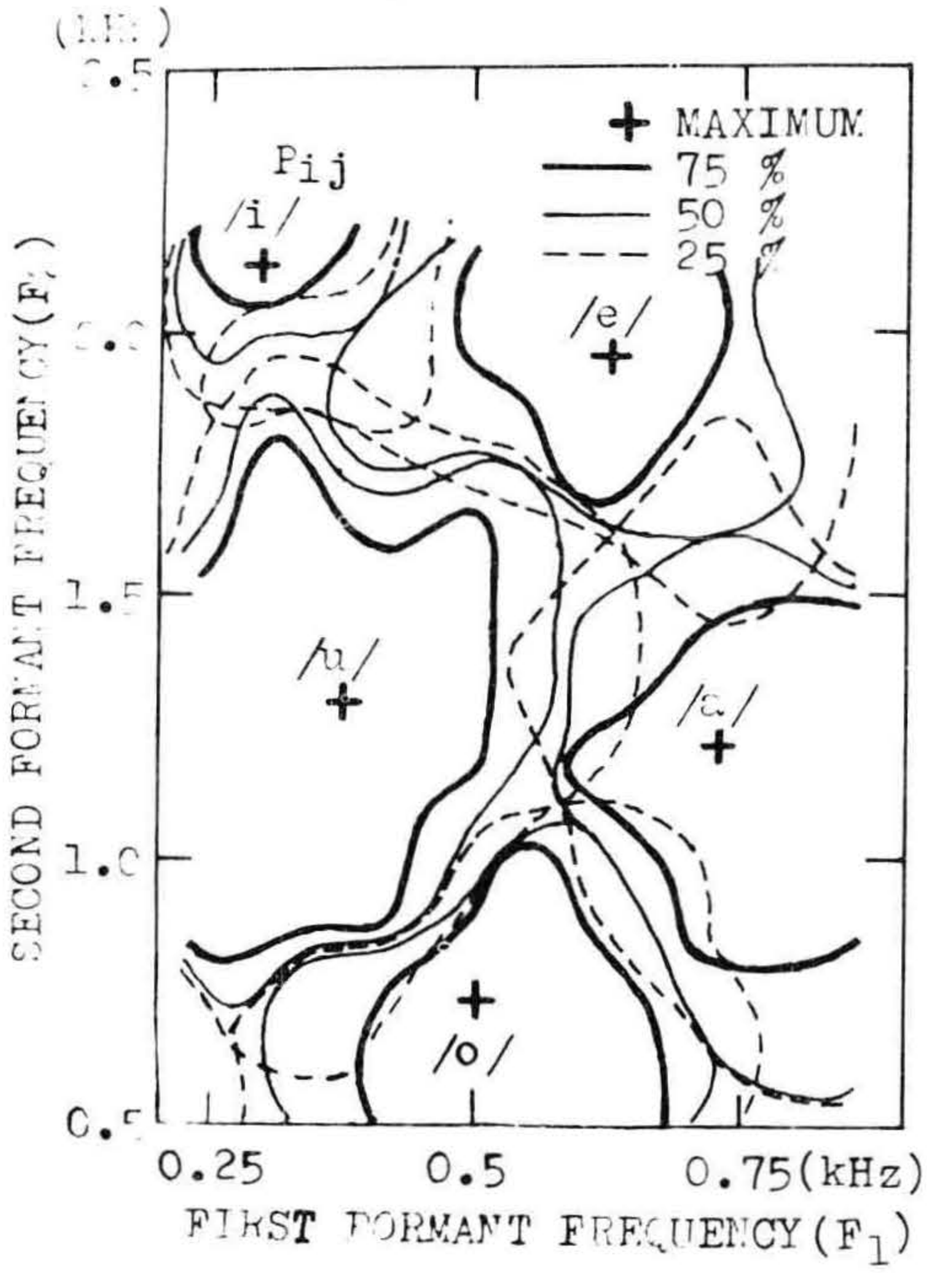


Figure 1. The probability of the identification of the synthetic vowel stimuli as one of five Japanese vowels (the average of five subjects).

pitch frequency:140Hz
third formant frequency:2.6kHz
fourth formant frequency:3.7kHz
duration:0.5sec.

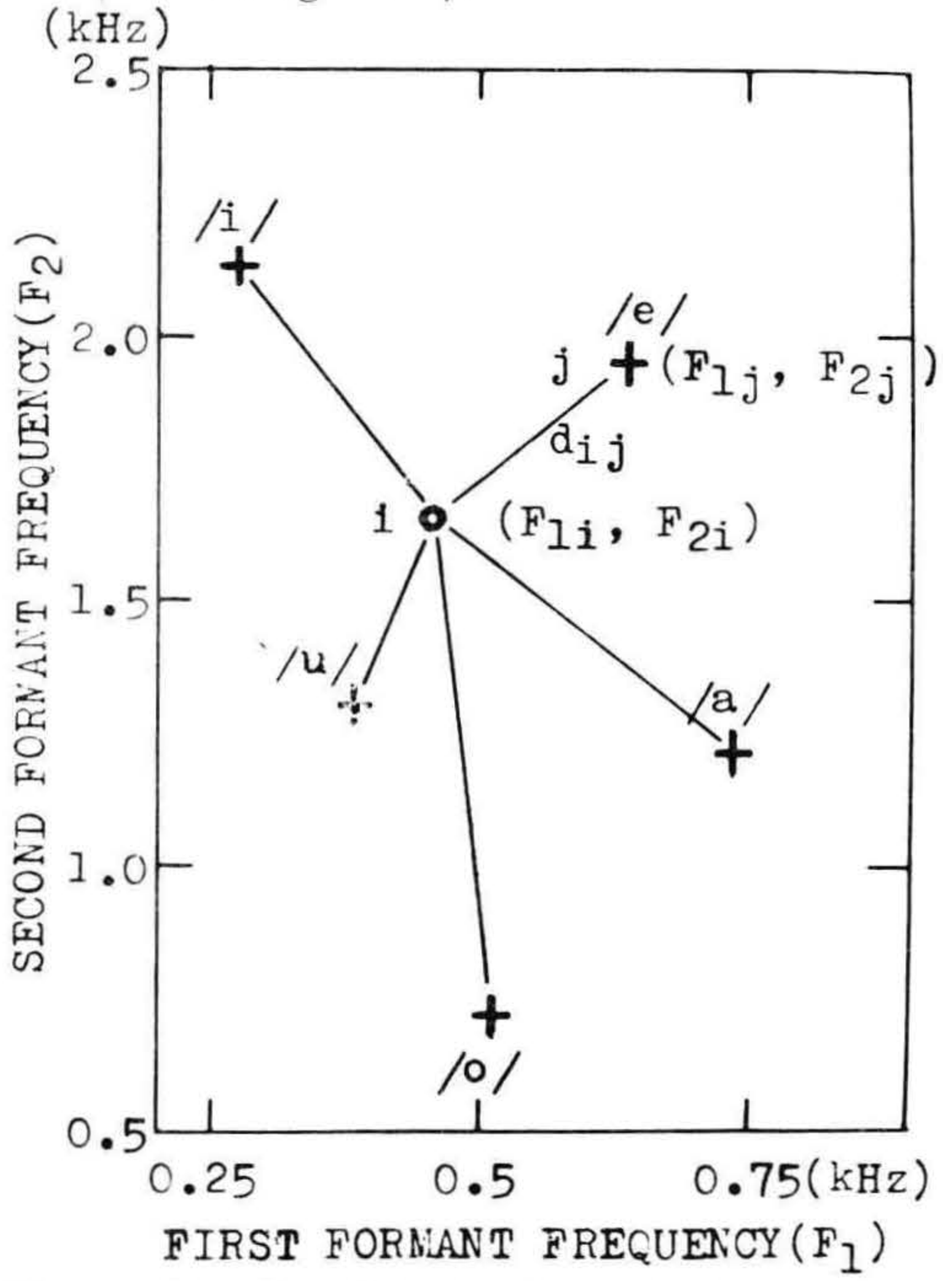


Figure 2. The illustration of the static model of the identification of the vowels. The degree of the resemblance is considered to have a relation to the distance from o to + (dij).

The process of the static perception can be formulated as follows;

$$P_{ij} = \left\{ 1 \,/\, 1 + (\alpha d_{ij})^{\beta} \right\} \,/\, \sum_{i=1}^{5} \left\{ 1 \,/\, 1 + (\alpha d_{ij})^{\beta} \right\},$$

where

$$d_{ij} = \left\{ (F_{1i} - F_{1j})^{2} + (F_{2i} - F_{2j})^{2} \right\}^{1/2} \text{(kHz)},$$

P_{ij} is the probability of the identification of the test vowel (i) as each one of five Japanese vowels (j), d_{ij} is the distance between i and j on the F1 - F2 plane with the linear scale of the frequency (refer Figure 2). With the scale factor α and the power β, it is estimated that $\alpha = 2 \times 10$ (kHz^{-1}) and $\beta = 4$.

Dynamic Model

When the test vowels were accompanied with the preceding vowels, the results of the identification tests are influenced by the difference in the phonetic quality between them (as shown in Figure 3 as an example). The process of the dynamic perception can be formulated as follows;

$$P'_{ij} = P_{ij} \left\{ 1 + \gamma e^{-t/\tau} (P_{ij} - P_{ij,\ pre.}) \right\},$$

where $P_{ij,\ pre.}$ is the probability of the identification of the preceding vowels as each one of five Japanese vowels by the static model, P_{ij} is the probability of the identification of the test vowels accompanied with the preceding vowels and t is the interval between two vowels in second (refer Figure 4). With the coefficient γ and the time constant τ, it is estimated that $\gamma = 0.8$ and $\tau = 0.5$.

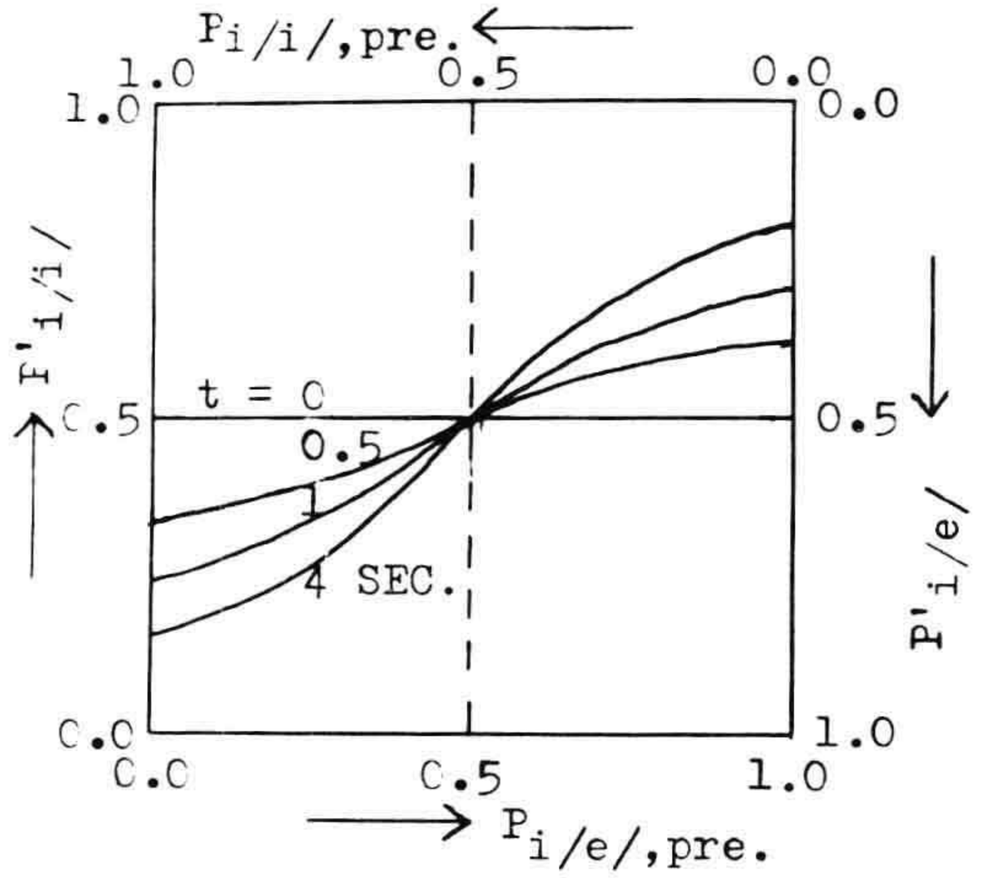


Figure 3. The change of the probability of the identification by the preceding vowels along the direction / i / to / e / on the F1 - F2 plane (the parameter is the time interval t).

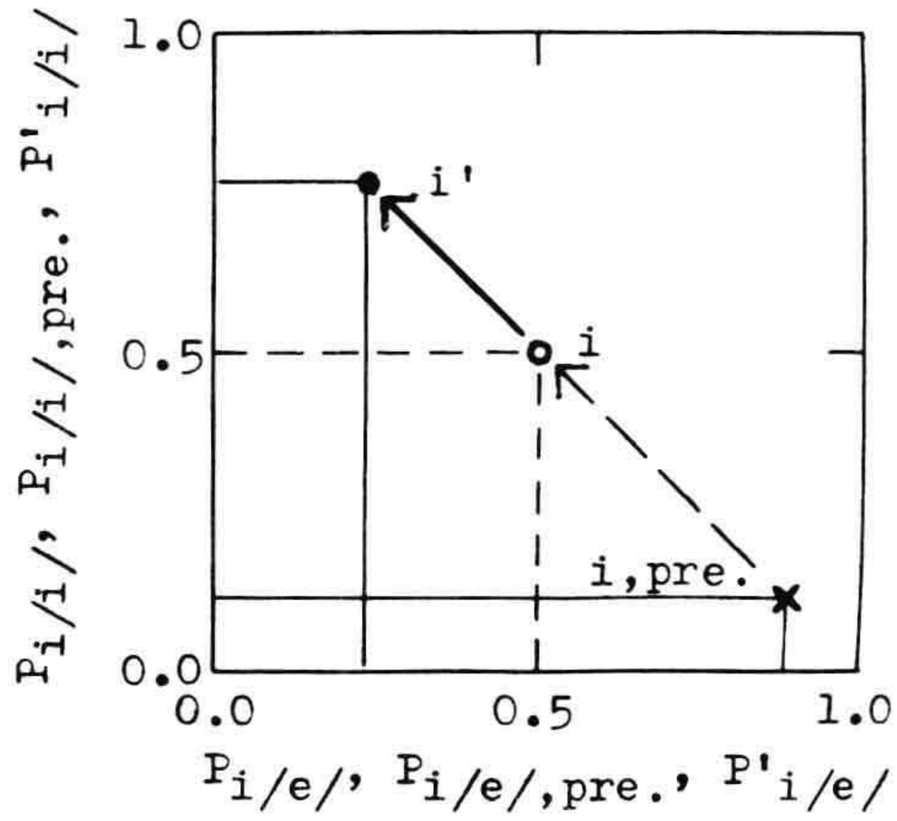


Figure 4. The illustration of the dynamic model of the identification of the test vowels as each one of five Japanese vowels.

Conclusion

It was examined that the dynamic model could be applied not only to the test vowels accompanied with the preceding vowels, but also to the ones accompanied with both the preceding and the following vowels and even to the ones accompanied with both the preceding and the following consonants which change the formants.

The formulations based on the mel scales instead of the linear scales of the formant frequencies and on the auditory space of the Japanese vowels (Reference 1) were also discussed.

Reference

1. Hideki Kasuya, et al.
"On the Auditory Space Model of Vowel Perception",
presented to the 6th ICA.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

The Role of Relative Spectrum Levels in Vowel Perception

Jan Ch. K-G. Lindqvist and Stefan Pauli
Research associates of the Department of Speech Communication, Royal Institute of Technology, Stockholm, Sweden

Introduction

There are two main directions in research on speech perception. One is to determine which acoustical parameters are measured by the auditory system and the other is to formulate a model that makes it possible to predict the perception. The speech stimuli can be described in many dimensions such as pole and zero frequencies of the transfer function and properties of the source function. There are, however, also other ways such as with properties depending on the time variation of the frequency spectrum. By systematically changing one parameter in a synthetic stimuli continuum and studying the responses in a forced choice classification test, it is possible to determine how the perception depends on that parameter. In the study of the perception of the vowels the interest has been concentrated on the effect of the formant frequencies, Fujimura 1967. In the experiments reported here we have studied the effect of amplitude changes in the vowel spectrum.

Stimulus specification

In our first experiment we used the front vowels [i] [y] [ʉ]. These vowels were synthesized with the OVE III series synthesizer. The stimuli were synthesized with a first formant frequency of 266 Hz and a varying frequency of $\sqrt{F2\ F3}$ in 12 steps. The mean value of F2 F3 is shown in Fig. 1. The F0 contour is the

same in all the stimuli. The spacing between F2 and F3 was five quarter-tones. The stimuli were played back through a filter that changed the relative level between the first formant and the higher formants by $\pm$ 12.5 dB. The total change in level is thus 25 dB. The overall level was corrected so that the vowels had about the same loudness.

Experimental procedure

Three test tapes were made. One with low level on the higher formants, one with high level on the higher formants, and one with both kinds of stimuli mixed in random order. Each stimulus was presented ten times, that is, there were 120 stimuli on the ordinary tape and 240 on the mixed tape. Eight subjects were used and they were all experienced in speech work. The stimuli were spaced with a time distance of 2.5 sec. In the first experiment the tape with the stimuli with only one type of filtration was used. In the second the mixed stimulus tape was used.

Results and discussion

The mean distribution curves for the responses of eight subjects are shown in Fig. 1. For four of the subjects the test was repeated four times with an interval of some days. The boundary between [i] and [y] was always very sharp and within one stimulus step. The boundary between [y] and [ʉ] was always more diffuse. This can be explained by the fact that the [ʉ] sound was not accepted as natural because these sounds are generally diphthongized in natural production. This diphthongization makes the difference between [y] and [ʉ] more clear. For the subjects who listened to the tapes several times the boundary shifted which indicates a sporadic variation in the classification. This sporadic boundary shift was about the same for one subject at different occasions as the variation among the different subjects.

From Fig. 1a we can see that the boundary between [i] and [y] is independent of the filtration. This is the case when we make the test with only one type of filtration at a time. The only difference is a greater uncertainty in the classification. For the boundary between [y] and [ʉ] the 50 % point is shifted in a direction

opposite to that expected. If we take into account the overall shape of the distribution curve it is better to describe this shift as an unchanged boundary determined with a greater uncertainty. In Fig. 1b the distribution curves are shown for the mixed stimuli. In this case we got a small but clear shift of the phoneme boundary.

From the first experiment we can draw the conclusion that we do not use any center of gravity or moment measurement when we make decision about isolated stationary vowels. Probably the auditory system measures the distribution of the spectral peaks instead. But we do not know if this is made in a broad-band manner or if we use the information of the frequency synchronous excitation of the hair cells along the basilar membrane to measure something that is more near the formant frequencies.

In the second experiment where the two kinds of filtered sounds were mixed we got a clear contrast effect, that is, the perception is influenced by the decision made on a previous stimulus. This contrast works in the amplitude dimension. If the decision mechanism only used the information of the frequency of the peaks in the spectrum we should only have contrast in the frequency dimension. Evidently we use the overall change in the distribution of spectral energy as an important parameter in the decision mechanism. In connected speech with constant source spectrum, the spectral level change is correlated with the change in the formant frequencies so this should be a useful complement to a tracking of the peaks in the spectrum. As the time between the stimuli were 2.5 sec the contrast must work on a central level. Perhaps the "difference" is a separate input to the decision mechanism when we make experiments with isolated sounds. One interesting experiment made by Fujimura (1967) with the same kind of stimuli as used here, where he changed both the separation and the mean frequency in a systematic way, shows that it is likely that we really use some kind of peak frequency measurement. On the other hand, two-formant synthesis shows that we can approximize the spectrum in a rather rough way without changes in the perception. If we look at the statistic distribution of nerve spikes in time and in intensity in the auditory nerve we have both information of something like zero-crossing frequency in a filter bank

and a more wide-band analysis. As we cannot reject either of these hypotheses we have better to assume, as a working model, that we use both types of information.

Coming experiments

In our coming experiments we are going to try to separate these two variables by comparing the sensitivity to level shifts on a three-formant synthesis with that of a two-formant synthesis. Of course, it is possible that we use fewer or other features when we listen to connected speech.

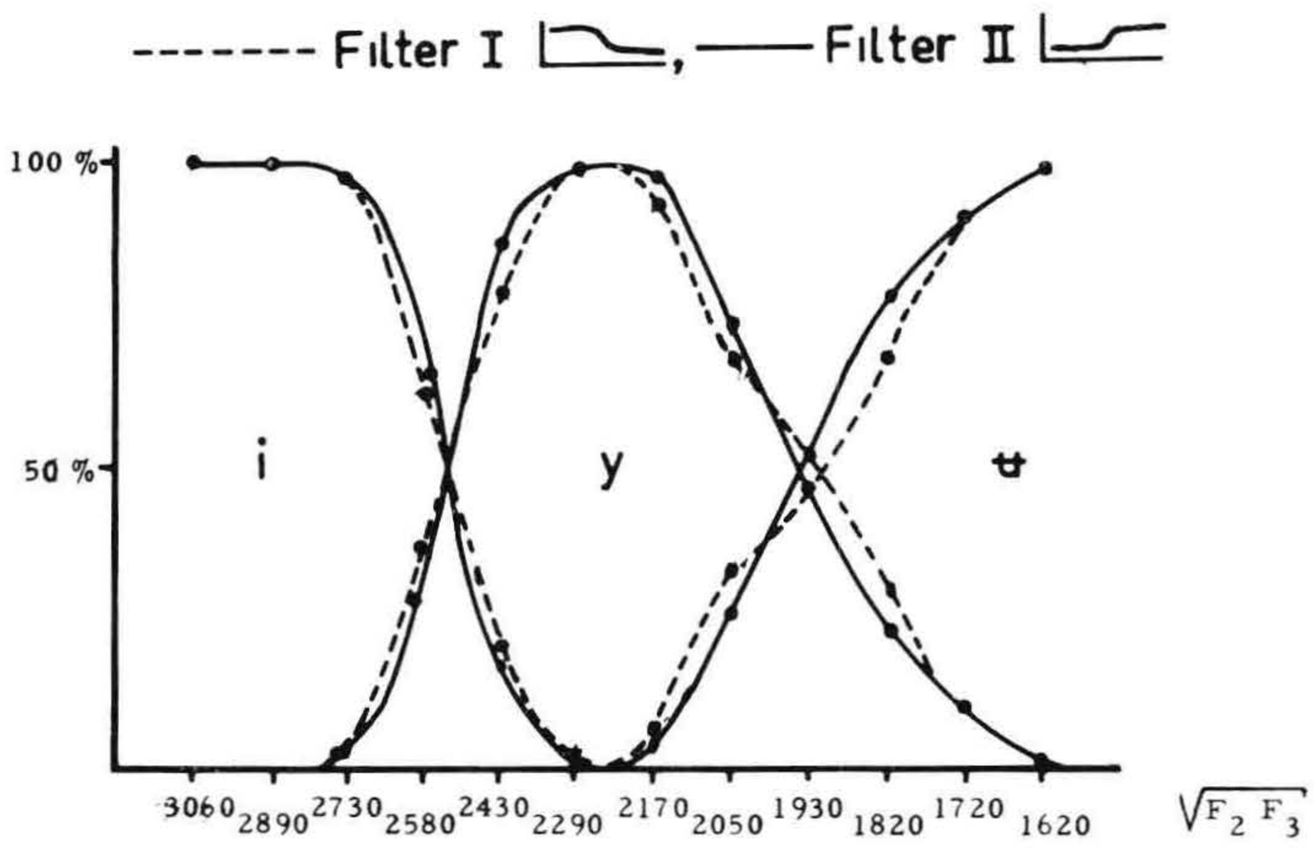


Fig. 1

Fig. 1. Forced choice classification of filtered vowels. a) Distribution curves of the responses when only one type of filtration was used in each test. b) Distribution curves when both types of filtrated stimuli were mixed.

Therefore we are also going to make synthesis of CVC syllables and see if we can make a simplified approximation of the acoustical features as, for instance, if we can approximate the formant transitions with only two formants.

References:

O. Fujimura: "On the Second Spectral Peak of Front Vowels ... A Perceptual Study of the Second and Third Formants", Language and Speech 10 (1967), pp. 181-193.

G. Fant and A. Risberg: "Auditory Matching of Vowels with Two Formant Synthetic Sounds", STL-QPSR (Stockholm), No. 4/1963, pp. 7-11.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-3-6

The Influence of Various Factors on the Identification and Discrimination of Synthetic Speech Sounds

Hiroya Fujisaki and Takako Kawashima
Engineering Research Institute
Faculty of Engineering
University of Tokyo
Tokyo, Japan

It has been reported that for a certain class of speech sounds, such as stop consonants, perception was found to be highly categorical, while for vowels, perception was found to be continuous. These findings seem to support the motor theory of speech perception. No further studies, however, seem to have been made so far of the origin of the characteristic difference, nor on the perception of other classes of speech sounds. In order to acquire deeper understanding of the mechanisms involved, we have studied the influence of various factors on the identification and the discrimination of synthetic vowel sounds and fricative consonants.

The Problem

Although published results indicate that the modes of perception are entirely different for vowels and for stop consonants, it remains to be investigated whether or not the difference results from differences in the acoustical properties of these two classes of sounds and in the experimental procedures adopted. Namely, stop consonants are relatively short in duration, transient in their character, and are always accompanied by a vowel that serves as a perceptual reference, while vowels have relatively long duration over which their acoustic characterictics are fairy steady, and can be produced in isolation so that they are not accompanied by any reference. These characteristics of vowels are also shared by fricative consonants. The individual influence of these properties can be studied by discrimination and identification tests using synthetic speech sounds.

Experimental Procedure

A cascade-type terminal-analog speech synthesizer was used to generate the test stimuli. The vowel section possesses controlled variables of four formant frequencies (F_1, F_2, F_3, and F_4) and the higher pole correction factor (K_{r4}), while the fricative section has controlled variables of a pole frequency (F_p) and a zero (F_z). These sections can be excited by a buzz or a noise source. The duration, the

intensity and the rise time constant of the excitation sources are also variable. The formant circuits can also be connected in parallel to simulate a parallel-type synthesizer. The accuracy of formant frequencies is better than 0.5 %.

For vowel discrimination tests, a set of 12 points was selected on a straight line in the log F_1 - log F_2 plane which included typical sounds for /i/ and /e/ of an average male speaker of Japanese as well as intermediate sounds, while F_3 and F_4 were held at 2750 Hz and 3850 Hz, respectively. A uniform spacing was adopted for F_1, since the results of preliminary tests indicated that it yielded approximately equal perceptual distance at both ends of the range. The values of F_1 and F_2 for these 12 stimuli were as listed in Table 1, and the band widths of these two formants were held approximately constant at 45 Hz and 90 Hz, respectively. The fundamental frequency of the buzz source was 130 Hz in most of the tests.

For discrimination tests of fricative consonants, a set of 11 stimuli was chosen over a range of pole frequency from 3000 Hz to 7810 Hz which included typical sounds for /ʃ/ and /s/. The ratio of the frequencies of the pole and the zero was 2. In this case, a constant ratio spacing was adopted between adjacent pole frequencies based on the results of preliminary tests. The Q of the pole was 5. The duration and the rise time constant of the stimuli were 100 msec and 20 msec, respectively. For fricative-vowel syllables, the vowel /e/ with a set of appropriate formant loci and rise time was added to each fricative sound, and the intensity ratio of the fricative and the vowel was kept at -6dB.

The ABX method was adopted for discrimination tests. Since any two stimuli that were more than two steps apart were almost always discriminated correctly, discrimination tests were performed only on pairs that were one and two steps apart. The interval between stimuli in an ABX triad was 1 sec, while successive triads were separated by 4 sec. For identification tests, the stimuli were recorded in random order and played back to subjects for forced judgment. Tapes for discrimination tests were also used for identification tests to examine context effects.

Six subjects, who were all adults with normal hearing and spoke standard Japanese (Tokyo dialect), participated in the listening tests. No pre-selection was made as to the performance of the subjects.

Results

Figure 1 illustrates the results of vowel discrimination tests with stimulus duration of 25, 50 and 100 msec. Each point represents the average of about 200 judgments by six subjects. The results of identification tests are shown for comparison. In spite of the averaging due to individual differences of phoneme boundary, the discriminability increases appreciably in the boundary region. Furthermore, perception is seen to be more nearly categorical for shorter stimulus duration. Minor variations in the discrimination curves can be partially ascribed to the harmonic structure of the spectrum of stimuli. In order to avoid this cause of variation, experiments have also been conducted using noise-excited vowels, and

the results, as in Fig. 2, also show higher discriminability in the boundary region.

For the purpose of comparison with the results in the literature, another test was conducted using stimuli generated by a parallel-type synthesizer. The lack of higher formants resulted in a shift of phoneme boundary, but the results, also shown in Fig. 2, reveal a similar tendency as observed in the foregoing experiments.

Finally, the effect of a reference sound was tested by adding a steady vowel /a/ of 50 msec duration immediately after each sound of a triad. This was considered to provide a constant environment for the test stimuli. The results are shown in Fig.3, which shows that perception becomes more categorical in this situation. The addition of the vowel /a/ is seen to cause a shift in the phoneme boundary, and, since the amount of this shift differs from individual to individual, the averaged results for two subjects with approximately the same boundary is shown in this figure.

Results of discrimination and identification tests of fricative consonants with and without a following vowel are shown in Fig. 4, which again reveals the same tendency as observed in the vowel tests.

Discussion

The apparent enhancement of discriminability at phoneme boundaries both in vowels and in fricative consonants, as evidenced by these experiments, can hardly be explained by referring to articulation. The explanation can be found partly in the twofold character of perception of speech and speech-like stimuli, which involves recognition of both discrete category and continuous quality, and partly in the task of discrimination in the ABX method, which requires short-time memory of three consecutive stimuli. In the discrimination test, identification also occurs upon reception of each sound, and memorizing the results of identification is evidently more reliable than memorizing the timbre of stimuli. Thus, the reliability of a subject's decision is bound to depend on whether stimuli A and B are identified as different or the same. Namely, in the former case, the task of discrimination becomes that of memorizing results of phoneme identification, while in the latter case the subject needs to have recourse to his less reliable memory of stimulus timbre for the final decision. The extent, to which a subject's overall performance becomes categorical, depends on the difference in reliability of the two memory channels involved , which in turn is seen to depend on various stimulus attributes and experimental procedures, such as duration, spectral features and environments of stimuli.

Conclusion

Results of discrimination tests on vowels and fricative consonants have revealed that perception of these sounds can be regarded, to a certain degree, as categorical which can hardly be explained by reference to articulation. Various factors have been shown to affect the discriminability, and an explanation is presented based on the twofold character of perception of speech sounds.

References 1. A. M. Liberman et al, J. exp. Psychol., 54, 358.
2. A. M. Liberman et al, Language and Speech, 4, 175.
3. D. B. Fry et al, Language and Speech, 5, 171.

Table 1. Frequencies of F_1 and F_2 of synthetic vowels used for discrimination tests.

F_1	(Hz)	310	330	350	370	390	410	430	450	470	490	510	530
F_2	(Hz)	2350	2298	2247	2197	2149	2101	2055	2009	1965	1921	1879	1839

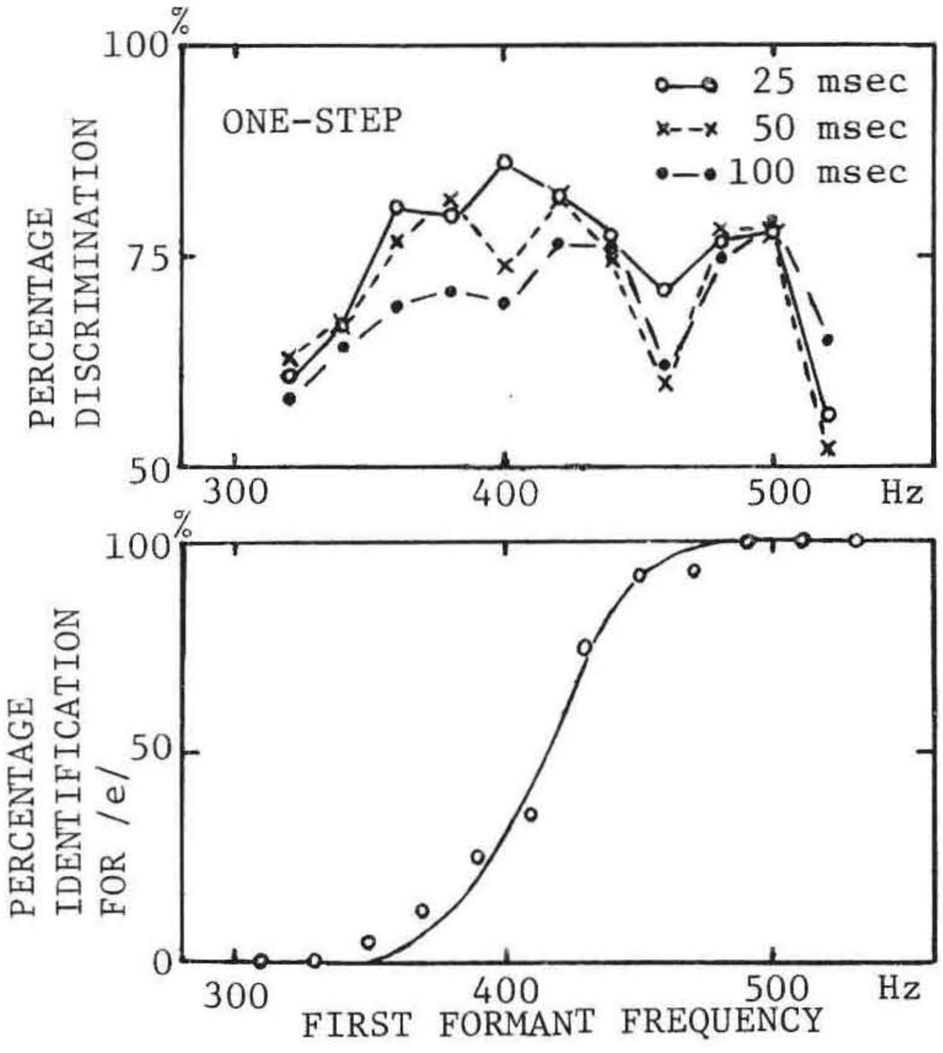


Fig. 1 Identification and discrimination of vowels /i/ and /e/. Effect of stimulus duration.

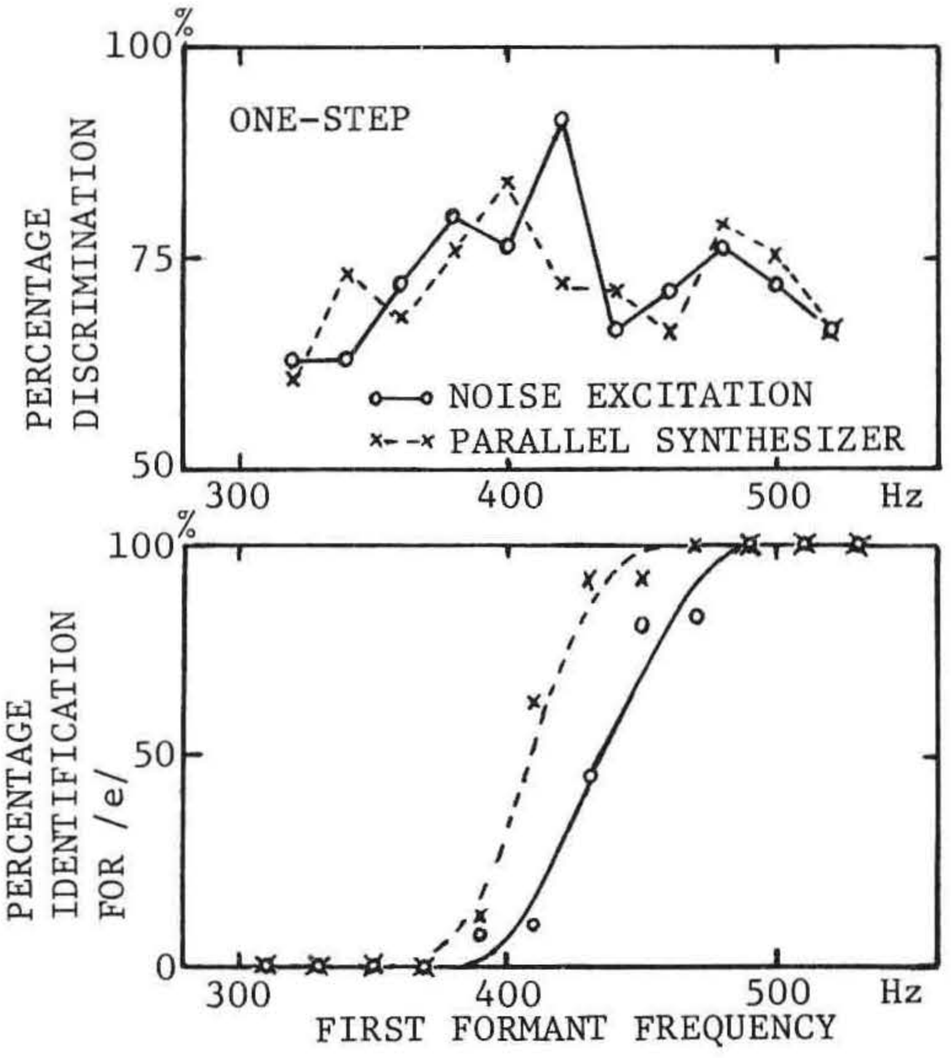


Fig. 2 Identification and discrimination of vowels. Effect of noise excitation and parallel synthesizer.

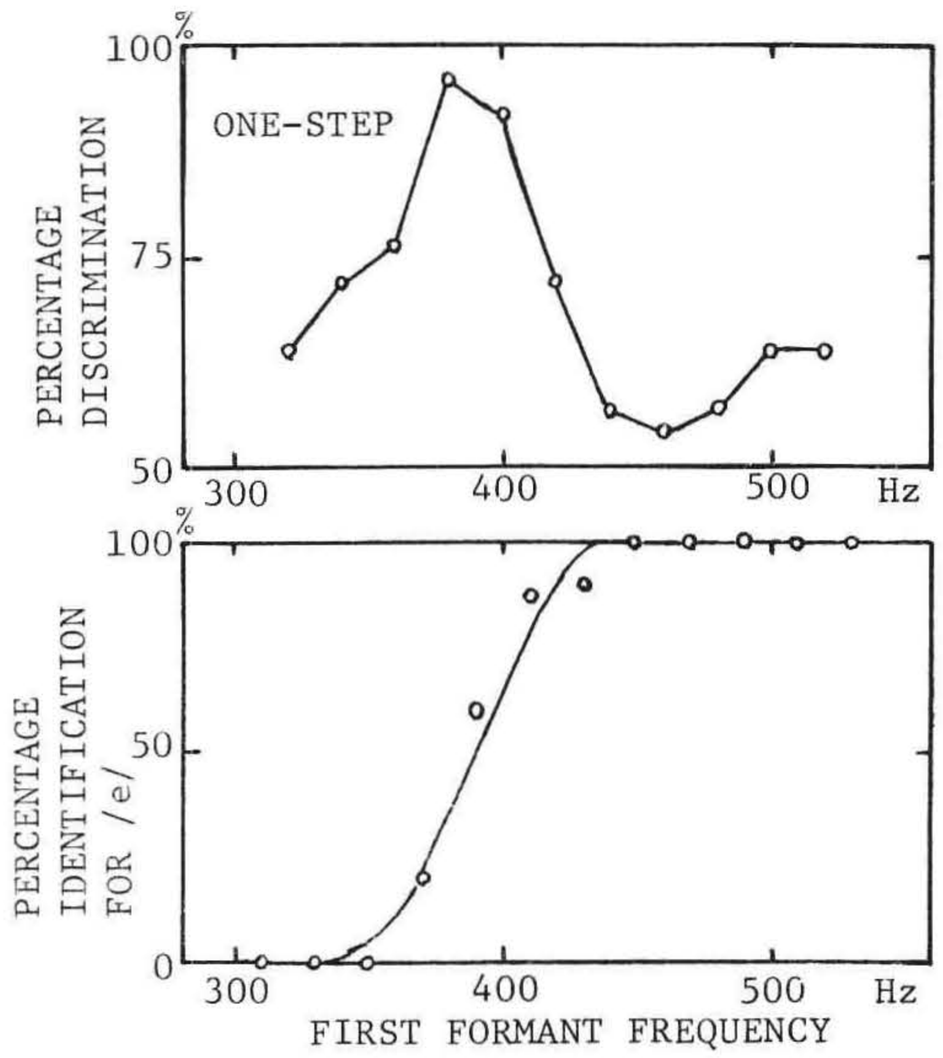


Fig. 3 Identification and discrimination of vowels. Effect of reference vowel /a/ following each stimulus.

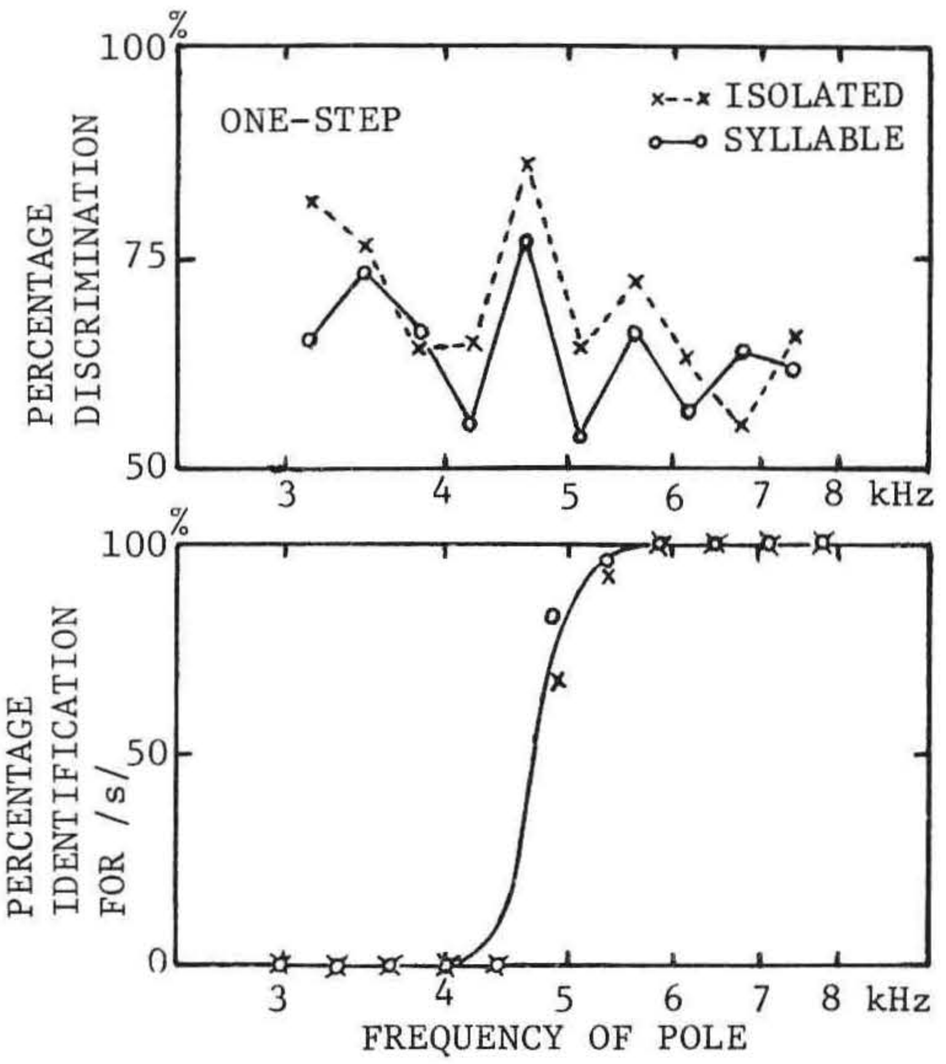


Fig. 4 Identification and discrimination of fricative consonants /ʃ/ and /s/. Effect of environment.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-3-7

Direction of transition as a perceptual parameter of time-varying stimuli

L.A.Chistovich
Pavlov Institute of Physiology
of Academy of Sciences of the USSR
Leningrad

Introduction

When the speech student says that the transitions are perceptualy important he may have in mind rather different things. He may simply refer to the well established fact that the transitional segments of speech flow convey useful information. In this case his statement concerns the distribution of information in speech stimulus and does not immediately concern the way of processing this information by human brain.

On the other hand the same statement may imply that some features of temporal alteration of the stimulus (its spectral shape, fundamental frequency, intensity) are detected by auditory system and are used as speech signal descriptors. Direction and rate of transitions have been suggested as the most probable general parameters which can characterize the temporal variations of quite different primary (carrier) parameters.

It is clear that perceptual importance of transitional segments in speech does not prove by itself the existence of specific transition-features-extracting mechanism. One way to prove its existence is to find some phonetic opposition which is based on the proper-

ties of transition and does not depend upon the absolute values of the carrier parameters. Two examples when this situation seems to exist will be presented in this paper.

1. Perception of F_o transition

The aim of the experiment was to test the perceptual importance of the fundamental frequency transition from consonant to vowel in CV syllable /1/. The syllable was constructed as a sequence of two "stationary" segments. Both of them were harmonic complexes. The consonant segment had two spectral peaks: the first one corresponding to F_o, the second one placed in the vicinity of 1.0 kHz. The spectrum of the vowel segment was shaped to resemble vowel a.

F_o at one of the segments was controlled by the experimentator. The subject had to vary F_o at the other segment and to try locate the boundary between ba and ma syllables.

The average boundary positions for four subjects are presented in Fig.1. The data can be approximated by a straight line representing the decision function. Its equation shows that only the ratio of F_o at consonant to F_o at vowel is relevant. The boundary corresponds to 10% increase of F_o from consonant to vowel. The data allow to postulate the existence of a threshold mechanism detecting the upward transition of fundamental frequency.

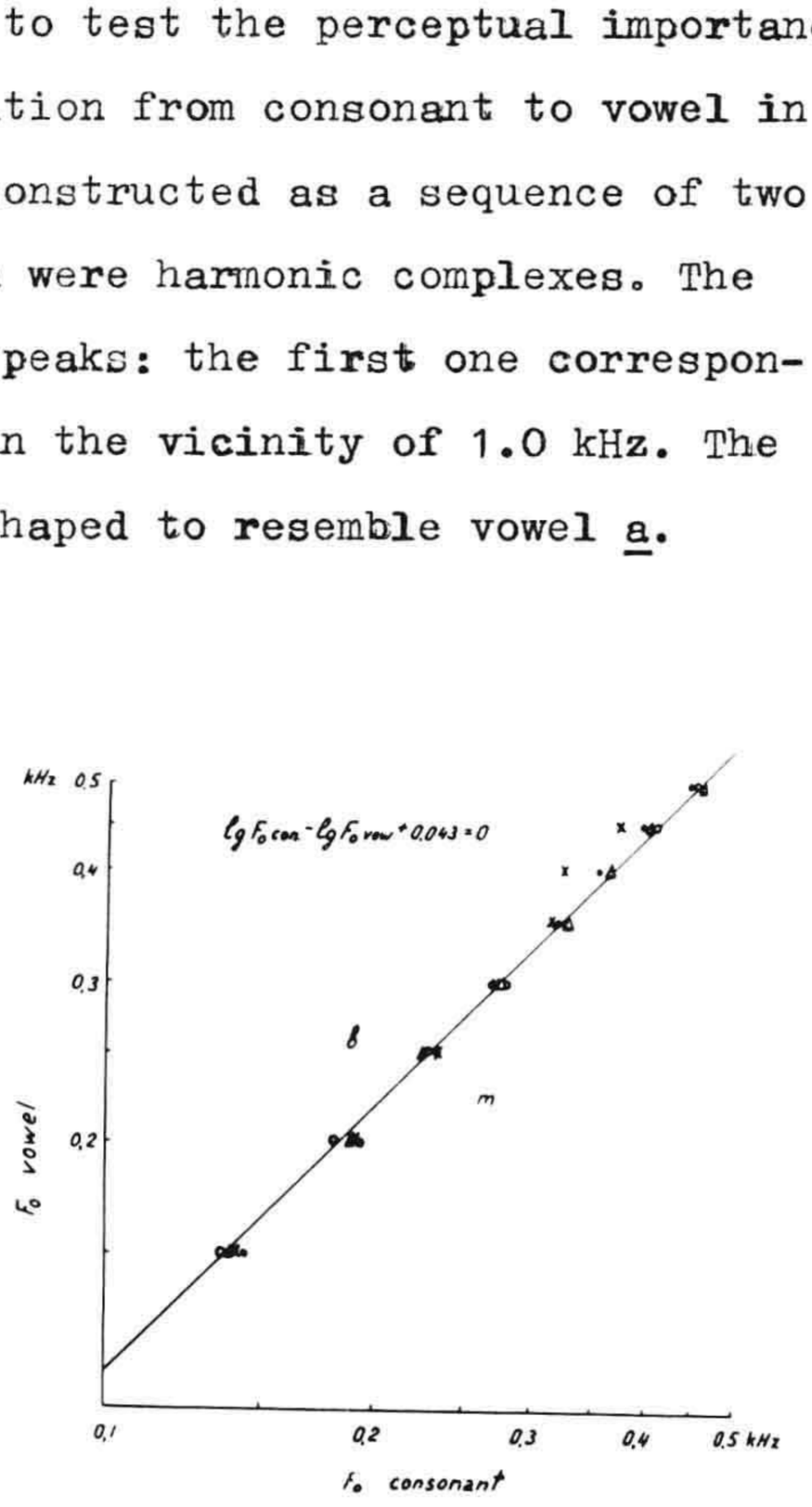


Fig.1

2. Perception of spectral maximum transition

The experiment previously described by Cooper et al. /2/ has

been reproduced with some modifications. The sequence of short (10-15 msec) burst of band-pass noise with variable midfrequency and vowel-like stationary segment was used as the test-stimulus. The vowel-like segment was a six harmonics complex, its spectrum envelope had a single maximum about 20 db high. This maximum was formed either by two adjacent harmonics or by a single harmonic. In the latter case it was possible to identify the frequency of the maximum with the frequency of this prominent harmonic. At low intensity level of noise burst the stimuli were perceived as pV or tV syllables.

The position of the spectral maximum of the vowel was controlled by the experimentator. The subject had to vary the midfrequency of noise and to try locate the p - t boundary.

The composite data for five subjects /3/ are presented in Fig.2. It shows the cumulative histograms of boundary locations (each subject made from 30 to 40 boundary adjustments) obtained in two different vowel conditions. Arrows in Fig.2 indicate the frequency position of two prominent harmonics in vowel spectra.

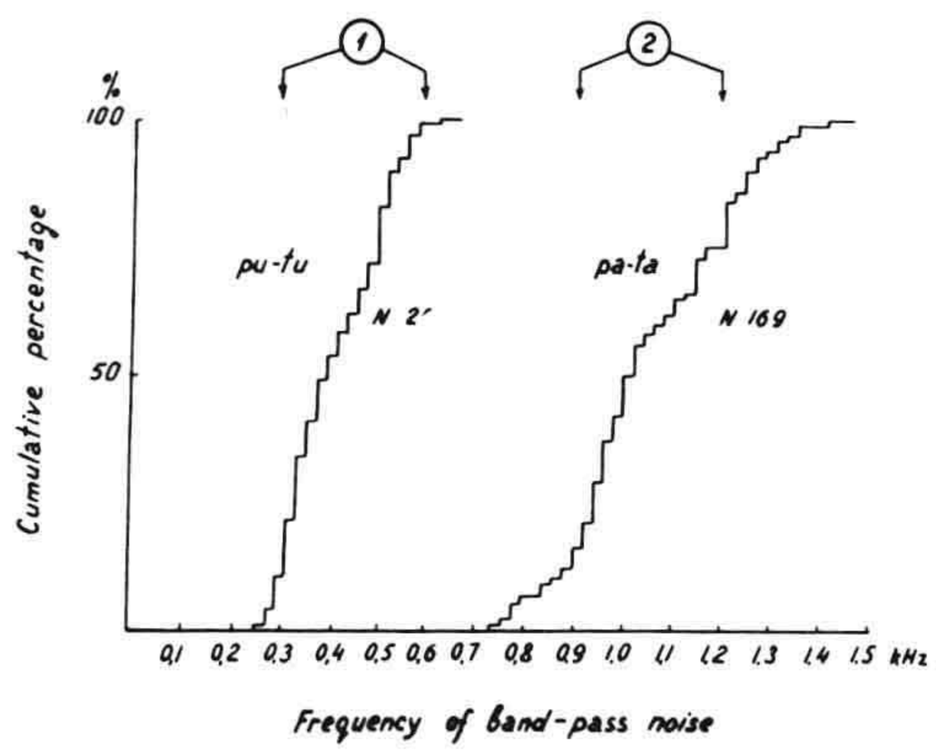


Fig.2

It can be seen that the p - t boundary distributions corresponding to different vowels do not overlap. This confirms the conclusion made previously by Cooper et al./2/ that both the noise and the vowel spectra are to be considered in order to make a decision about the place of articulation of stop consonant.

Fig.2 also shows that the main part of each distribution falls

between the frequencies of the most prominent harmonics of the vowel. This makes to think that the average boundary position coincides or is rather close to the frequency of the maximum in the spectrum envelope of the vowel.

To test this hypothesis another experiment was made /4/. The data for two subjects are presented in Fig.3. Each point represents the mean of 20 boundary adjustments.

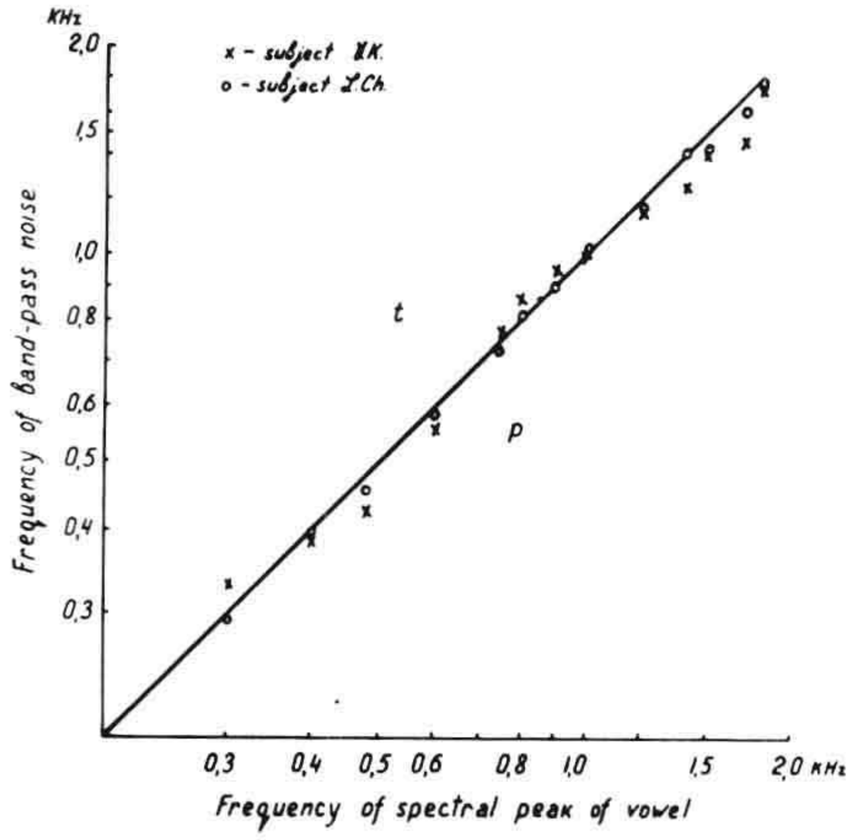


Fig.3

Fig.3 shows the average boundary position as a function of the frequency of the most intensive harmonic representing the spectral maximum of the vowel.

The coincidence of the boundary with the frequency of the vowel spectral peak implies that p - t discrimination is based on the direction of spectral maximum transition from consonant to vowel. The falling transition corresponds to t, the rising - to p.

Recent data on acoustic correlates of the place of articulation for stop consonants /5/ suggest that the direction of transition must be an effective cue also in natural speech recognition.

References:

1. Chistovich L.: Proc.of the 6th Intern.Congr.of Phonetic Sciences,Prague,1967. 2. Cooper F.S. et al.: J.Acoust.Soc.Am.,24,597-606, 1952. 3. Reitblat L.E.: Perception of direction of the spectral maximum transition in synthetic syllable. Graduate thesis,1968. 4.Kozhevnikov V. a.L.Chistovich: Auditory measurement of the formant frequency,1968. 5. Stevens K.N.: 1967. Conference Speech Communication and Processing. Boston, Mass., 177-184.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-3-8

Distinctive Feature: A Context-Free Element

Sadanand Singh

Associate Professor of Speech

Howard University

Washington, D. C. 20001

Introduction

Speech perception is a process of detecting distinctive units of acoustic signals structured in the context of a language. The phoneme is one convenient set to contain such units. The phoneme, however, is an interpretive unit. The interpretation is dependent on, among other things, a network of distinctive acoustic signals that separate one phoneme from another. The distinctive acoustic signal is a physical phenomenon and the perception of the phoneme psychological. In order to determine the relevance of a distinctive feature to its phonemic correlates, perceptual tests were conducted in the present study to investigate the importance of the voiced/voiceless contrast in English.

Among the distinctive features voicing is the most widely used by both the descriptive and experimental phoneticians. There may be controversy regarding its physiological correlates; but the acoustic descriptions of voiced/voiceless is more or less a universal one. Voicing has been characterized by the presence or the absence of 'glottal buzz' (Lisker and Abramson, 1964). The 'larynx is free either to generate voice or to remain mute' (Lisker and Abramson, 1967). Halle, Hughes, and Radley (1957) maintain that the difference between these two classes of stops lies in the fact that in the production of the voiceless stops more pressure is built up behind the closure (fortis or tense) than in the production of the voiced (lenis or lax). These functions may result in the presence or absence of voicing, aspiration, and strength. The three parameters signal the same distinction of voiced/voiceless. In some recent experiments, it has been conveniently diagnosed that the acoustic parameter of time distinguishes the contrast effectively (Singh, 1966; Barrs, 1966; Grimm, 1967; and Lisker and Abramson, 1967).

Procedure

In the present study 16 English consonants were used in the prevocalic position.

Fourteen included seven voiced/voiceless pairs;/p, b; t, d; k, g; f, v; ð, θ ; s, z; and tʃ , dʒ./ and two were /m/ and /n/. Twenty-five native subjects of American English listened to the recorded syllables following the psychophysical method of ABX. The experiment was designed to include only those choices in an ABX response frame that were one feature different from a stimulus. The determination of one feature difference was based on the Miller and Nicely (1955) and Halle (1964) systems of classification of consonants. For example, for the stimulus phoneme /p/, /b/ and /t/ were minimally different using both the Halle and Miller and Nicely systems; /k/ using the Halle system only; and /f/ using the Miller and Nicely system only. A similar description is applicable to the inclusion of response choices /p/, /d/, /g/ and /v/ to the stimulus /b/. The inclusion of /d/, /p/, and /θ/ as response choices to the stimulus /t/ satisfied the requirement of minimal difference of both systems of classification. However, /t/ and /k/, according to Halle, differ because of the features gravity and diffuseness. According to Miller and Nicely they are minimally different by a place of articulation. A similar description, again, is applicable to the voiced counterpart stimulus /d/ and response choices /t/, /b/, /g/, and /ð/. The phoneme /t/ was minimally different from /k/ using the Miller and Nicely system and /p/ using Halle. Again, a similar description applies to the voiced counterpart stimulum /g/ and response choices /k/, /d/, and /b/. The phonemes /f/ and /θ/ and /f/ and /p/ were minimally different using the Miller and Nicely system and /f/ and /s/ using Halle. In the same manner other response choices were included following either one or both of the two systems of classification.

The 16 stimulus consonants and the response choices to each of them can be seen in Table 1. Each of the response consonants appearing with a given stimulus was paired with the others an equal number of times and presented in a counterbalanced order. For example, when the stimulus was /p/ and 12 ABX and AXB frames were formulated pairing b, t, k, and f with each other. There were 164 such stimulus-response frames for the 16 stimulus consonants.

Table 1.

Rank Order of Responses	Stimuli /p/	/b/	Stimuli /t/	/d/	Stimuli /k/	/g/	Stimuli /m/	/n/	Stimuli /f/	/v/	Stimuli /θ/	/ð/	Stimuli /s/	/z/	Stimuli /tʃ/	/dʒ/
1.	b	p	d	t	g	k	n	m	v	f	ð	θ	z	s	dʒ	tʃ
	47%	40%	42%	37%	58%	48%	80%	80%	47%	34%	36%	33%	39%	35%	43%	50%
2.	t	d	p	b	t	d	b	d	θ	ð	f	v	ʃ	ʒ	ʃ	ʒ
	25%	27%	23%	28%	22%	28%	20%	20%	28%	28%	23%	20%	35%	34%	40%	48%
3.	k	g	k	g	p	b	p	b	s	z	θ	ð	θ	ð		
	18%	19%	20%	20%	20%	24%	14%	20%	18%	17%	14%	16%	17%	2%		
4.	f	v	θ	ð			s	z	t	d	f	v				
	10%	14%	15%	15%			11%	18%	16%	16%	12%	15%				
5.									tʃ	dʒ						
									7%	14%						

The subjects marked their judgments by circling numbers in each frame that appeared on the response sheet. For example, if they heard the frame which contained /pa/, /ta/, /ka/, the question asked was whether /pa/ was more similar to /ta/ or to /ka/. If to /ta/, circle number (2) and if to /ka/, circle number (3). Use of this procedure assures avoidance of the confusing effects of the graphemic similarity between the sounds.

Results

The results of this study are summarized in Table 1. The stimuli are presented horizontally and responses vertically. The responses are presented in the rank-order of frequency of judgments. For example, to the stimulus /p/ the most frequently marked response was /b/; second to it was /t/, etc. Since each response consonant related to a stimulus was paired with the others, there was an equal number of appearances of each response consonant. Therefore, a hypothesis was formulated that there is no difference between the frequency of the response consonants that are one feature different from a stimulus. The hypothesis was confirmed in more than 50% of the comparisons. The results of an independent $\underline{X^2}$ test for each frame showed that in more than 50% of the individual comparisons (e.g. /p/→/b/-/t/) showed significant difference at > .1% level.

There is a striking consistency in the manner in which the responses rank. For example, the rank-order of responses to the stimulus phoneme /p/ is parallel to that of the stimulus phoneme /b/, carrying only the distinction that distinguishes /p/ from /b/: voicing. Thus it may be stated that:

1. /b/, /t/, /k/, and /f/ : /p/ = /p/, /d/, /g/, and /v/ : /b/.
2. /d/, /p/, /k/, and /θ/ : /t/ = /t/, /b/, /g/, and /ð/ : /d/.
3. /g/, /t/, and /p/ : /k/ = /k/, /d/, and /b/ : /g/.
4. /v/, /θ/, /p/, and /s/ : /f/ = /f/, /ð/, /b/, and /z/ : /v/.
5. /ð/, /f/, /s/, /t/, and /tʃ/ : /θ/ = /θ/, /v/, /z/, /d/, and /dʒ/ : /ð/.
6. /z/, /ʃ/, /θ/, and /f/ : /s/ = /s/, /ʒ/, /ð/, and /v/ : /z/.
7. /dʒ/, /ʃ/, and /θ/ : /tʃ/ = /tʃ/, /ʒ/, and /ð/ : /dʒ/.

The order in which they are presented is based on frequency of response, ranging from the consonants bearing the most frequent response to those bearing the least frequent response. It is evident from the table that the degree to which the frequency declined in each case was not constant. In the manner of stating results of this experiment it may also be stated that:

/n/ and /b/ : /m/ = /m/ and /d/ : /n/.

One important hypothesis tested in this experiment was the relevance of a given feature in different phonemic contexts. Tested was the hypothesis that the difference of a given feature, for example voicing, in any one phonemic contrast situation,

would not yield significantly different frequencies of responses in other situations of phonemic contrast by the same feature. Comparison of all scores associated with any one of the features voicing, place, frication, or duration, using the Miller and Nicely system of classification did not show significant differences. Similarly, the scores associated with any one of the features voicing, gravity, diffuseness, continuant, or stridency did not show any difference using the Halle system of classification. The test of significance was $\underline{X}^2$ and level of significance at $>.1\%$. These comparisons were all limited to the intra-feature category. The only scores compared were the ones associated with different phoneme pairs marked by the same distinctive feature. For example, scores associated with /p-b/ were compared with /t-d/, /k-g/, /f-v/, /θ-ð/, /s-z/, and /tʃ-dʒ/. Based on the findings that they were not significantly different it was concluded that voicing was an independent feature of phonemic contexts.

Conclusion

The results of this experiment showed the consistency with which voicing maintained the interconsonantal spacing. The findings that for a stimulus, for example /p/, the 1st, 2nd, 3rd, and 4th most frequent responses were /b/, /t/, /k/, and /f/ enable one to predict that to the voiced counterpart, the phoneme /b/, the 1st, 2nd, 3rd, and 4th most frequent responses would be /p/, /d/, /g/, and /v/. This indicates that the /p/-/b/ contrast is consistent not only with /b/-/p/ but also with /t/-/d/, /k/-/g/, and /f/-/v/, disregarding the interacting influence of any other feature.

References

Barrs, J.T., Identification of voiceless initial fricatives versus modification of their duration, Linguistics: An International Review, 21, 5-23, (1966).

Grimm, W.A., Perception of segments of English-spoken consonant-vowel syllables, J. Acoust. Soc. Amer., 40, 1454-1461, (1966).

Halle, M., On the basis of phonology, Structure of Language: Readings in the Philosophy of Language, eds. J. Foder and J.Katz, Englewood Cliffs, N.J., (1964).

Halle, M., Hughes, G.W., and Radley, J.-P.A., Acoustic properties of stop consonants, J. Acoust. Soc. Amer., 26, 107-116, (1957).

Lisker, L. and Abramson, A.S., A cross-language study of voicing in initial stops: acoustic measurements, Word, 20, 384-422, (1964).

Lisker, L. and Abramson, A.S., Some effects of context on voice onset time in English stops, Language and Speech, 10, 1-28, (1967).

Miller, G.A. and Nicely, P.E., An analysis of perceptual confusion among some English consonants, J. Acoust. Soc. Amer., 27, 338-352, (1955).

Singh, S., Cross-language study of perceptual confusion of plosive phonemes in two conditions of distortion, J. Acoust. Soc. Amer., 40, 635-656, (1966).

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

Is the consonant-vowel amplitude ratio an index of talker intelligibility?

Michael S. Weiss and Arthur S. House

Department of Audiology and Speech Sciences, Purdue University, Lafayette, Indiana

---

Research findings have demonstrated that raising the talker's level of vocal effort decreases the intelligibility of the speech materials (Pickett, 1956) and decreases the consonant-vowel amplitude ratio in the speech signal (Fairbanks and Miron, 1957). Furthermore, it has been suggested that, everything else being equal, a talker's relative intelligibility may be largely determined by his characteristic consonant-vowel amplitude ratio (House et al., 1965; Williams et al., 1966). This report deals with an investigation aimed at demonstrating the degree to which the relative intelligibility of a group of talkers was related to the consonant-vowel amplitude characteristics of the talkers.

The speech materials used in the investigation were the test lists described by House et al. (1965), as amended (House and Guttman, 1966). They are six 50-word lists of monosyllabic English words in consonant-vowel-consonant form, generated in a manner which provides for the formation of 50 ensembles of six words each. The listening task is a closed-response set one in which the listener

selects his response from a six-word ensemble where the words are characterized by the same vocalic nucleus (and orthographic representation) and differ only in their initial or final consonant.

Twelve talkers, all adult males, each recorded four forms of the test materials. Forms and word orders were systematically assigned, assuming that a range of consonant-vowel amplitude ratios would be produced by the group of talkers, and that four forms per talker would provide an adequate basis for checking the equality of test forms. All talkers had a preliminary training period and significant dialectic idiosyncracies were excluded. The talkers were coached to maintain a constant (conversational) level of vocal effort without instrumental monitoring, to maintain a consistent intonation pattern across the carrier phrases and test words, and to produce weakly released or unreleased final consonants. The recordings were made on high-fidelity equipment under quiet conditions.

Each recorded list of words was analyzed by obtaining the rms voltage output of a graphic level recorder used in a mode that provided clearest differentiation between the vocalic and consonantal portions of the utterances, and consonant-vowel amplitude ratios (in dB) were determined (except for a few words that were unmeasureable, and for words using voiced stop consonants which were omitted from the study).

Eight adult males listened binaurally to the recorded lists. The lists were presented at a level of approximately 70 dB SPL under the phones and white noise was mixed with them to provide speech-to-noise ratios of -4 and +2 dB. Presentation of voices, test forms, speech-to-noise ratios, etc., was varied systematically.

Results

The obtained consonant-vowel amplitude ratios ranged from approximately -9 dB to almost -15 dB. In general, the ratio for a given talker was consistent over four test lists, but for a few talkers the mean consonant-vowel ratio at one speech-to-noise condition differed by more than one dB from that in the other condition. On the average, however, the consonant-vowel ratios for the tests presented in the two speech-to-noise conditions were not significantly different.

As expected, the mean intelligibility scores for the two speech-to-noise conditions were about 50 and 70 per cent. The correlations between intelligibility scores and consonant-vowel ratios, however, were not high enough for predictive purposes. Unexpectedly, the rank-order of talkers by intelligibility was considerably different at the two speech-to-noise conditions in the study, and examination of various classes of sounds found no standard categories whose predictive value materially exceeded that of the over-all test lists.

Examination of the responses to the various forms of the test lists showed that each form provided highly similar levels of response. In both listening conditions and when responses are pooled over conditions, the test forms were essentially equivalent.

These general findings fail to support the hypothesis that a talker's consonant-vowel ratio is a major determinant of his word intelligibility, as suggested earlier by House et al., and Williams et al.

References

Fairbanks, G. and Miron, M. S. Effects of vocal effort upon the consonant-vowel ratio within the syllable. J. acoust. Soc. Amer., 29, 1957, 621-626.

House, A. S. et al. Articulation testing methods: consonantal differentiation with a closed-response set. J. acoust. Soc. Amer., 37, 1965, 158-166.

House, A. S. and Guttman, N. Personal communication, 1966.

Pickett, J. M. Effects of vocal force on the intelligibility of speech sounds. J. acoust. Soc. Amer., 28, 1956, 902-905.

Williams, C. E. et al. Intelligibility test methods and procedures for the evaluation of speech communication systems. Report No. ESD-TR-66-677, Decision Sciences Laboratory, ESD, AFSC, USAF, L. G. Hanscom Field, Bedford, Massachusetts.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-3-10

On the Experimental Evaluation of the Interior Redundancy in Speech

Werner K. Endres
Research Institute of the
"Fernmeldetechnisches Zentralamt der Deutschen Bundespost"
Darmstadt, W-Germany

Introduction

It is known that the duration of certain phonemes if spoken at normal speed is longer than required for reliable recognition by man.

In an earlier investigation by the author [1] it was shown that it should be theoretically possible to reduce the time required to transmit normal speech by approximately 50 percent. This conclusion was based on the assumption that "redundant phoneme segments" can be eliminated without effecting recognition. "Redundant phoneme segments" are defined as those parts of phoneme segments which exceed 30 ms in duration and which possess time-frequency spectra which do not change appreciably during the phoneme utterance.

In the investigations reported here the postulated temporal (or time) redundancy of phonemes has been analysed experimentally. The purpose of the investigations was to determine whether the theoretically postulated redundant phoneme segments, if eliminated, do or do not effect the logatom intelligibility and hence represent truly redundant speech segments.

Method for Determination of Temporal Redundancy

To achieve reproducible results and to separate linguistically

and semantically introduced redundancy from the acoustical time-redundancy only meaningless words - so-called logatoms - have been used for the investigations. These test-words consist of random phoneme sequences and are selected to be free of context associations.

Logatom word lists were spoken at normal speech. From the tape-recordings visible speech spectrograms were made. Analysis of these permitted the identification of redundant segments, that is of segments for which the frequency spectrum remains relatively unchanged over periods exceeding 30 ms. Corresponding segments were than eliminated from the tape recording as described below.

Experimental Details

Five experiments were conducted. For each experiment 150 logatoms were accordingly prepared. Each set of processed logatoms was evaluated by two crews of 5 resp. 4 listeners. Evaluation of the measurements was achieved following the directions given by Schneider [3].

The nature of the 5 experiments and the results obtained are summarized in table I. Reduction of a vowel in a logatom was achieved by cutting out sections of a certain duration from the tape recording. This was done in three steps (B, C, D), beginning in the middle of the phoneme and proceeding to the neighbouring phonemes. The remaining phonemes of the logatom have not been reduced.

Reduction of long consonants was achieved in the same manner (Experiments No. 1 and 2 in table 1).

In experiment No. 3a all phonemes and all silent parts have been reduced to a minimum value absolutely necessary for recognition. As a result, the duration of the logatoms was still only about 35 % to 50 % of the original duration. Therefore in experiments 3 b and c the reduced logatoms were repeated two, resp. three times, thus re-

introducing a new, but different temporal redundancy.

The reduced logatoms of experiment No. 3 have been expanded in time by a factor of 1,26 in experiment No. 4 and normally spoken logatoms have been compressed in time by a factor of 0,56 in experiment No. 5, using the "expansion-compression method" of Springer [4].

Table I. Forms of Experiments and Results.

Experiment No.	Form of Experiment	Quantitative Reduction %	Logatom Intelligibility %
1 and 2			
A	No reduction		92
B	One phoneme reduced	45	90
C	One phoneme reduced	68	70
D	One phoneme reduced	91	24
3 a	Each logatom totally reduced	57	64
b	2-fold repetition of each logatom		70
c	3-fold repetition of each logatom		73
4	Time expansion	125,6	66
5	Time compression	55,8	74

Results (See table I)

The test results can be interpreted as follows:

Experiments 1 and 2: Vowels and consonants of long duration can be reduced to about 55 % of the original duration without loss in intelligibility. By a further reduction to about 30 % one obtains still acceptable intelligibility rates.

Original Information Lost.

Original Information Lost.

we attempt to find some regularities of Japanese words and discuss the possibility of their application to the word recognition. Our investigation is carried out using about 6,000 most frequently used Japanese words.*4 The accumulated frequency of occurrence of these words is about 86 per cent.

Segmental Units of Japanese Words

There are three types of segmental units of words, thst is, a phoneme, a mora and a syllable. There are 24 phonemes in Japanese.*5

V(vowels):/i/,/e/,/a/,/o/,/u/. Y(semi-vowels):/j/,/w/.

C(consonants):/p/,/t/,/k/,/b/,/d/,/g/,/m/,/n/,/ŋ/,/r/,/h/,/s/,/z/, /'/ (laryngeal phoneme).

P_m:/N/ (syllabic nasal), /Q/ (choked sound).

These phonemes are connected systematically with one another to make morae which can be connected almost freely except for P_m. There are 110 morae:

M: CV --- /ta/,/'o/,/zu/, etc. CYV --- /kja/,/'jo/,/mjo/, etc.

P_m: ----- /'N/,/Q/.

Syllables are: M, MP_m, ML, where L denotes the elongation of vowels. Syllables can also connect almost freely with one another, but they are large in the number (about 2,000). The number of phonemes is very small, but their combination is restricted strongly. Therefore, a mora is suitable as the unit in the investigation of regularity of the word composition in view of their small number and freedom in connection.

Phonological Characteristics of Japanese Words

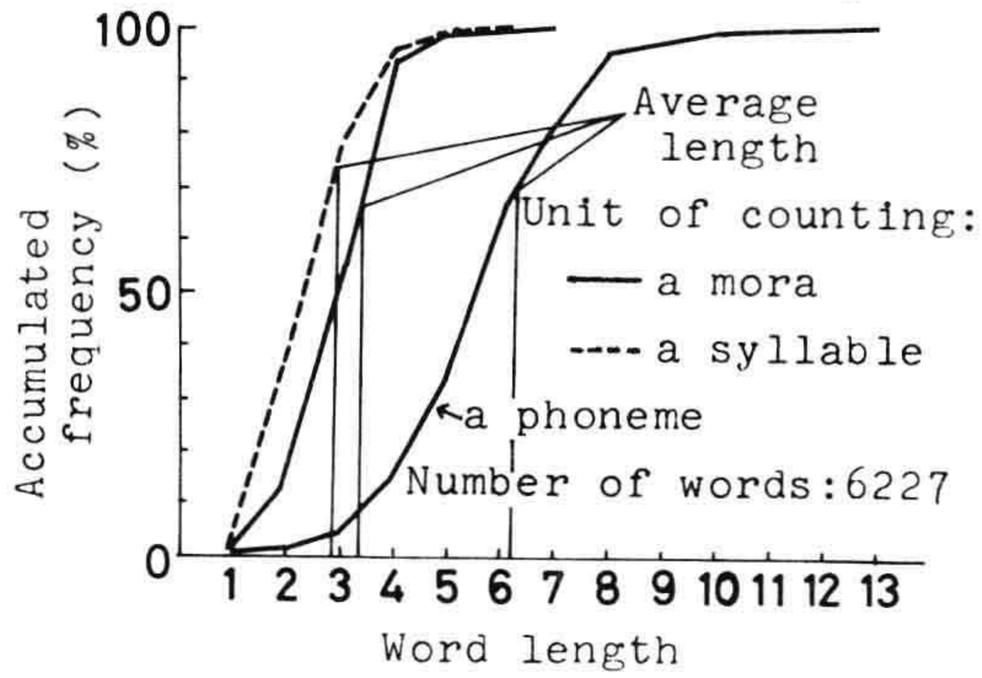
1) Word Length The accumulated frequency distribution of word length is shown in Fig.2, which was counted using a phoneme, a mora and a syllable as the unit respectively. The average lengths are also shown in Fig.2.

2) Redundancy The number of all possible sequences of morae N_m is

$$\sum_{k=1}^{r} 110^k = 1.46\times10^8 > N_m > \sum_{k=1}^{r} 108^k = 1.36\times10^8,$$

where r is the length of the longest word actually used. The accumulated distribution reaches 93.5 per cent at 4 morae in Fig. 2, so the number is estimated approximately (putting r=4) at $N_m = 1.40 \times 10^8$. As the vocabulary of a usual dictionary is about 50,000, the redundancy of mora sequence becomes approximately

Fig.2 Accumulated frequency distribution of word length



$$R_m = 1 - \frac{\log 5\times10^4}{\log 1.4\times10^8} = 1 - 0.577 = 0.423 .$$

3) Distribution of Morae in Respect of Place in Words

The frequency distribution of occurrence of morae was calculated in which each mora was counted separately in respect of the place of each mora in words. Elongated vowels /L/, a syllabic nasal /N/ and a choked sound /Q/ do not occur in the top of words. The mora /ru/ scarecely occurs in the top but mainly does in the end of words.

4) Digram of Morae in Words

The frequencies of one mora following another, so called digram frequencies, have been measured. Some of results are as follows. See Fig.3.

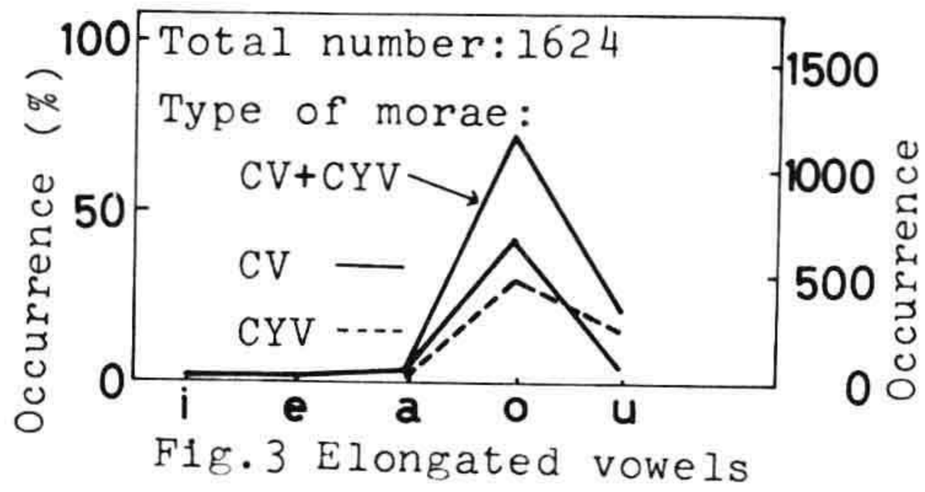


Fig.3 Elongated vowels

Number of investigated words:6227.

Number of all items of the digram:

12544 (= 112×112)

Number of items each of which has the frequency:
- zero :9759 (77.8 % of all items)
- one :1035 (8.3 % of all items)
- more than one:1750 (13.9 % of all items)

Redundancy of mora-pairs: $R_p = 1 - \frac{\log 2785}{\log 12544} = 1 - 0.84 = 0.16$.

Original Information Lost.

Original Information Lost.

This system is capable of evaluating features precisely and correctly. This paper is on the pre-recognizer and the paper on the post-recognizer will be submitted later.

Feature Extraction and Recognition

The samples used are human voices of 100 male, all of whom do not have any special training. As a feature extractor, a formant extractor using zeros of S.S.B. waves is used. The features used here are 1st formant, 2nd formant, numbers of syllables, energy variation of sound and etc., and they are shown in the Table I along side with threshold level for coding.

Features extracted	Threshold for coding	Values of sample"8"	Code
1.Interruption or non-interruption of energy		Interruption	1
2.Average frequency of the 1st formant of the 1st vowel	680Hz	690Hz>680Hz	1
3.Average frequency of the 2nd formant of the 1st vowel	1,650Hz	1,200Hz<1,650Hz	0
4.Average frequency of the 2nd formant of the 2nd vowel		1,850Hz>1,650Hz	1
5.Time derivative of the 1st formant	+200Hz/s		0
6. "	-200Hz/s	|-100 Hz/s| <|-200Hz/s|	0
7.Time derivative of the 2nd formant	+400Hz/s	+4,800Hz/s >400Hz/s	1
8. "	-400Hz/s		0
9.Number of syllables	3	3=3	1
10.Existence of the 1st formant	Range 450Hz 70Hz	None	0

Table I

When an example of "8"(hachi)is taken, the formant movement as an output of the formant extractor is shown in Fig.I. The coding example of this digit is shown in the Table I. As seen in Table I, the standard code of "8" is 1101001010. When it being differently voiced, it becomes 1101001110. In case it is possible to map this error code into the standard code, the correct decision should be obtained. Now V represents the feature space, which is the space where the binary codes of 10 tuples obtained with 10 threshold is spanned. U represents the decision space which is a sub-space of V

spanned with 10 standard codes. (The standard code is an average code obtained statistically from the code of each pattern class of spoken digit.)

The procedure in which the voiced spoken digits are correctly recognized corresponds with the procedure in which the standard space U is corresponded and this mapping procedure is the very procedure of error correcting codes. Though each bit being apt to show the same error rate in the error correcting codes of communication system, in this case the same error rate is not recognized. Thus, bits with much error become to have priority in the process of being corrected. For this purpose, suitable weight has to be added to each bit of the codes for error correction. W denotes the weight value of code a and is expressed as follows;

$$W(a) = \sum_{i=1}^{10} p_i a_i$$

where $a = (a_1, a_2, a_3, \cdots, a_{10})$

p_i = error probability of i th bit

Now $\dim V=10$, $\dim U=7$ being obtained, the feature space can be separated into 10-7 cosets. When the code with the least weight is chosen for a coset leader, the bits with the larger error rate are corrected and the codes among , which amount to be $2^{10}=1,024$ are all mapped into U which amount to $2^7=128$ codes. The space U includes some codes which are not standard codes. When they are devided into 10 code-groups according to each value of error rate, the system gives 98.3% of correct decision per 10 digits of about 100 males.

When values of error rate for decision are clipped by some threshold level, say 0.1, the pattern of codes of which error rate is below 0.1 is not recognized. The result of our experiment showed that 20.7% of spoken digits has left unclassified and the correct decision of 79.2% was resulted. In this case, however, the measured value expressing to which pattern the error rate from each code of standard pattern is the closest, is given as the imput information to

Original Information Lost.

Original Information Lost.

By the duality principles of linear programs, the linear programming problem (1) may be transformed immediately to the following dual problem

$$\psi(A, B)=\min\left\{g(r+s) \,\middle|\, Au-Bv+r-s=0,\quad eu=1,\quad fv=1,\quad u\geq 0,\ v\geq 0,\ r\geq 0,\ s\geq 0\right\}, \tag{2}$$

where r and s are n-dimensional column vectors.

Utilizing the fact that the solution $\psi(A, B)=\theta(A, B)$ converges to zero when A and B are linearly inseparable and that a pattern corresponds to a variable in (2), it is possible to determine sequentially a set of linear functions which separates the two disjoint pattern sets $A=(A_1, \ldots, A_p)$ and $B=(B_1, \ldots, B_q)$ by the following procedure:[2)]

1° Solve the linear program (2) for B and A_1, the first pattern of A. If the solution (A_1, B) 0, add the second pattern A_2, increasing the number of the variables of (2) by one, and optimize further the linear program if necessary. Thus continue to add $A_3, A_4, \ldots, A_j$ as long as $\psi_j(A_j^*, B) > 0$, $A_j^*=(A_1, \ldots, A_j)$.

2° If $\psi_j=0$, reject the pattern A_j and restore the simplex tableau to the state before A_j is added, and again continue to add A_{j+1}, etc. When the last pattern A_p is added, the first pass through the set A is end and a linear discriminant function $c(A^1)$ to separate A^1 from B is determined as the dual solution of (2), where A^1 denotes the set of patterns not rejected in the first pass.

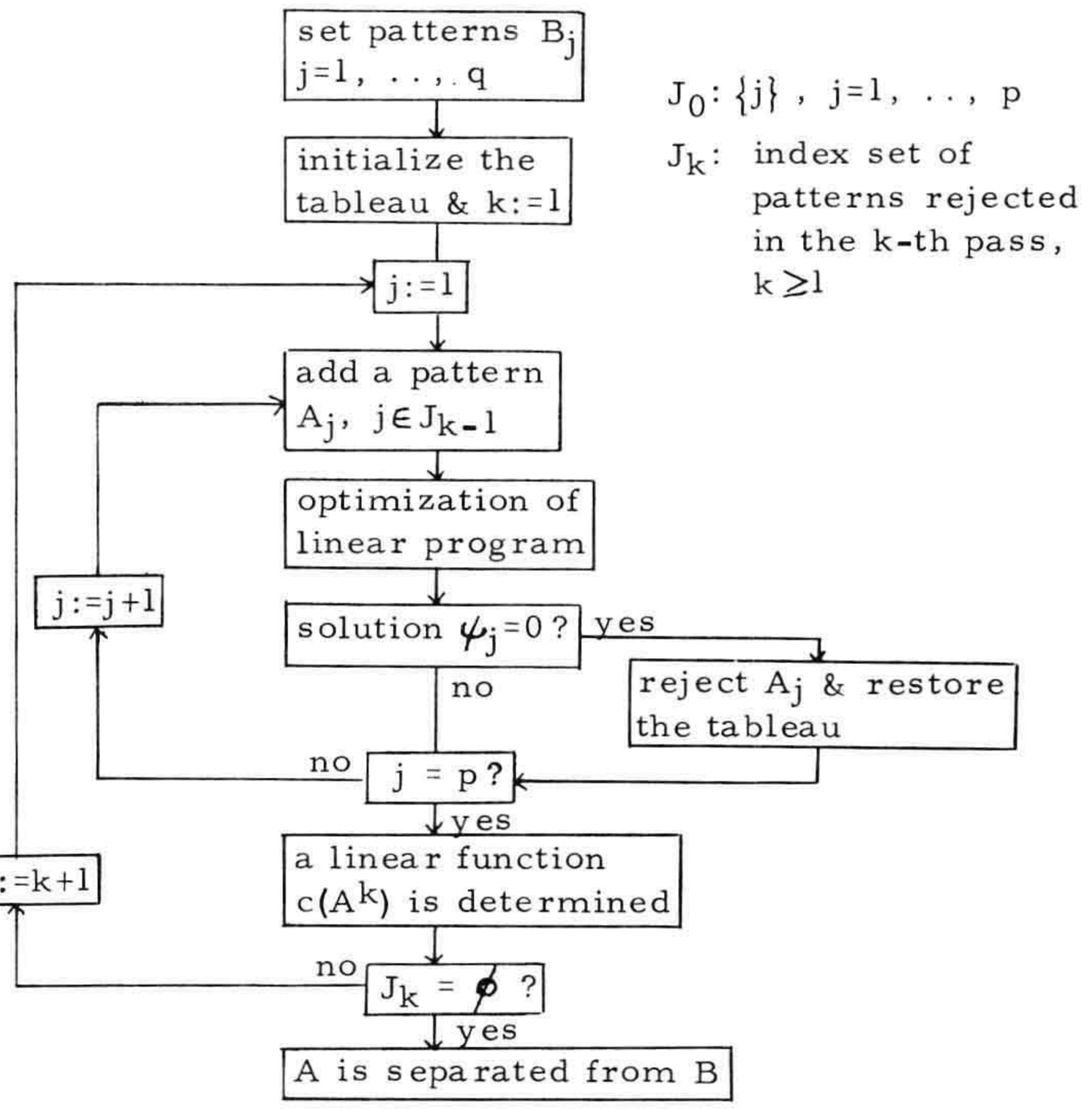


Fig. 1 Pattern separation procedure using linear programming

3° For the patterns rejected in the first pass apply the above process and determine $c(A^2)$. Such a process is repeated until no pattern is rejected through the pass and the entire procedure terminates.

4° If all patterns added in the k-th pass are rejected, exchange the role of A and B, and determine $c(B^1)$, $c(B^2)$, etc. for $A-(A^1, \ldots, A^{k-1})$ by the similar process as in 1°, 2° and 3°.

The process described in 1°, 2° and 3° is illustrated in Fig. 1.

Computer Program

When the magnitude of patterns has no meaning, patterns should be separated by the linear boundary passing through the origin of the measurement space, and therefore β in the problem (1) should be equal to - . Accordingly, the dual problem (2) must be modified slightly as follows:

$$\phi(A, B)=\min\left\{g(r+s) \mid Au-Bv+r-s=0,\quad eu+fv=1,\ u\geq 0,\ v\geq 0,\ r\geq 0,\ s\geq 0\right\}. \quad (3)$$

The computer program coded for the NEAC L3 is based on the formulation (3) because the object of separation is the speech represented by the spectrum components. In this program the revised simplex method with product form inverse is used for its simplicity in rejecting variables. The rule to add patterns is altered to improve the efficiency of computation : After the initial feasible solution is obtained, the linear separability of each pattern yet to be added is examined, at every iteration, for the current dual solution vector given as the simplex multiplier, and patterns found separable are added immediately.

Results of Preliminary Experiment

As a preliminary experiment on spoken word recognition, linear discriminant functions of ten Japanese spoken digits have been computed. The speech signal is analyzed, after high-emphasis, by a bank of seven channel bandpass filters contiguously covering 100 to 6000Hz, the outputs of which are envelope detected by fullwave rectifiers and lowpass filters of 10Hz cutoff. The output voltages are then sampled at every 45msec, multiplexed and converted to nine bits codes, which are then written on the digital tape of the computer through the real-time channel. One word is sampled at eight points, and represented as the 7x8=56-dimensional vector, as shown in Fig. 2. Each of ten digits is uttered once by fifteen male speakers.

The linear discriminant functions to separate the pattern sets of each class of ten digits from other patterns then were computed, setting p=15 and q=135. An example of linear functions obtained is shown in Fig. 3 as the 56-dimensional solution vector, which corresponds to the same class as shown in Fig. 2. Each set of patterns was separated from others by only one linear function. The number of iterations required for optimizing each linear program ranged from 78 to 191. The computation time per iteration takes about ten seconds.

Conclusion

A procedure to separate patterns by multiple linear discriminants has been

Original Information Lost.

Original Information Lost.

spectral pattern (P(t)). Since it can be considered that the selected sections are in the same processes of pronunciations respectively, the standardized pattern can be obtained. This can be regared as a kind of normalization in time direction. The original measurement is expressed as $X(f,t)=(X(1),...,X(t),..,X(m))$, where $X(t)=(x_1(t),x_2(t),..,x_n(t))$, each $x_i(t)$ being the output of the i-th bandpass filter of n channels at time t. For P(t), we took the following value;

$$P(t)=\sum_{i=1}^{n}\left[\frac{x_i(t)}{\sqrt{\sum_{i=1}^{n}(x_i(t))^2}}-\frac{x_i(t+\Delta t)}{\sqrt{\sum_{i=1}^{n}(x_i(t+\Delta t))^2}}\right]^2. \qquad (1)$$

In Fig.1, value of P(t) of Japanese monosyllable (JA) is shown. On the analysis of the figures of P(t) of many monosyllables, some typical properties can be noticed. It is known that spoken syllable with a consonant is divided into a consonant part, a transition part and a stable vowel part. A sharp peak of P(t) can be seen just before the beginning of the transition part and value of P(t) becomes low at a stable interval of a vowel part and it decreases prom the peak to the stable point in the transition interval gradually. The normalization of a spectral pattern to time component is performed by picking up fixed numbers(k) of sections from the original measurement. It is desirable to take the mean of the neighboring three sections to the corresponding section and to normalize the amplitudes of mean measurements. Then the normalized measurement can be expressed as $Z=(Z(1),Z(2),..,Z(k))$. Here $Z(k)=(z_1(k),z_2(t),...,z_n(k))$, where

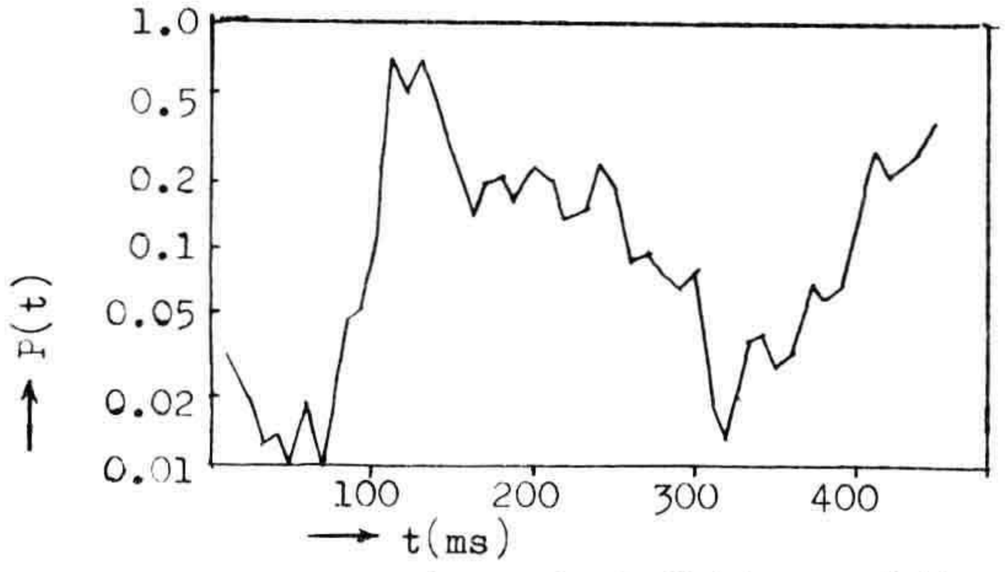


Fig.1 Deviation of Spectral Pattern of the Sound "JA"

$$z_i(k)=\frac{(x_i(k-1)+x_i(k)+x_i(k+1))}{\sqrt{\sum_{i=1}^{n}(x_i(k-1)+x_i(k)+x_i(k+1))^2}}. \qquad (2)$$

We recognize the pattern regarding it as an kxn dimensional vector.

The procedure to determine the decision function to be used in

this paper is the technique to obtain the optimal recognition code which is assigned to match the relations in the measurement space to those in the code space best and to determine linear decision fuctions to classify patterns according to the recognition code[3,4].

Japanese monosyllables (JA),(JU),(JO),(ZA),(JI),(ZU),(ZE),(ZO) were used for experiments. For the former three, 4 different persons pronounced in three different ways, that is, long, short and with an accent and 1 person pronounced in the normal way. For the latter five, 5 persons pronounced in the normal way. The outputs of 18 channels of a filter bank were sampled at 100 times per sec.

In the first experiment three sections were taken for the normalized pattern. The rules to determine the sections to pick up are as follows: (1) The first section (s1) is referred to the information of a consonant part. It is taken as the center section among sections where P(t) exceeds 0.3. (2) The second section (s2) is referred to the information of a stable vowel part. It is taken as the center sections among sections where P(t) is less than 0.07. (3) The third section (s3) is referred to the transition part. It is taken as the center section between s1 and s2. In the second experiment two more sections were added to the first three sections to pick up more informations from a consonant part and a transition part. The fourth section (s4) was taken at the second section preceding to s1 and the fifth section (s5) was taken at the center between s2 and s3.

Fig.2 shows value of variance of distance from the mean vector in

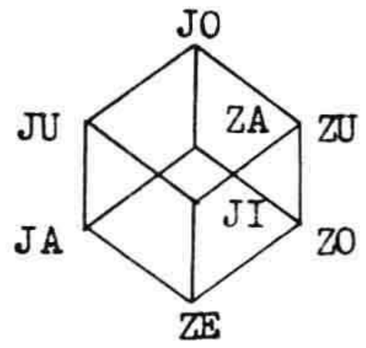
Experiment \ Pick up Section	s1	s2	s3	s4	s5	Mean Variance
(1) 3 Sections from P(t)	0.204	0.147	0.164			0.172
(2) 5 Sections from P(t)	0.204	0.147	0.164	0.222	0.199	0.187
(3) 3 Sections from Spectral Pattern	0.185	0.144	0.161			0.163

Fig. 2 Variance of the distance from the mean vector.

the first and second experiment and they are compared with that of the case that three sections were taken from the spectral pattern, according as a consonant, transition and stable vowel part, as an ideal case. The optimum recognition code and the obtained connecting graph for the first experiment are shown in Fig.3 together with the number of iterations until all input data are recognized beyond the threshold in the adaptive procedure. Performance of the recognition resulted with no error per 64 samples. For the second experiment the optimal recognition code was equal to the first one and the performance of the recognition was almost the same with no error.

The variance for the first experiment is approximately the same as that of the case that three points were taken from the spectral pattern. Since the latter value may be considered as the result of the variation of different speakers, and the recognition was carried out successully, it may be said that the normalization was made adequately. The difference dependent on the number of normalized measurements was not revealed in these experiments. However, it will become apparent when the number of phonemes increases.

	T1	T2	T3
JA	1	0	0
JU	1	0	1
JO	0	0	1
ZA	0	0	0
ZI	1	1	1
ZU	0	1	1
ZE	1	0	0
ZO	0	1	0
I	33	18	20



I: Number of iterations until optimal weights of decision functions are obtained

Fig. 3 Recognition code and connecting graph in experiment 1

The authers wish to thank Dr.H.Shinkawa, Dr.S.Ohshima and Dr.Y. Nakagome of KDD Labratory and Dr.H.Enomoto of Tokyo Institute of Technology for their advice and encouragement.

References

1. P.Denes, M.V.Mathews, J.A.coust.Soc.Am.,32 (1960) 1450.
2. J.H.King,Jr., C.J.Tunis, IBM Journal, Jan. (1966) 65.
3. H.Enomoto, S.Inoue, A.Kurematsu, ICMCI,Tokyo (1964) I-11-2.
4. H.Enomoto, S.Inoue, A.Kurematsu, Trans.of Electronics and Communication Engineers of Japan,51-C (1968) 113.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-4-6

Discrimination of Japanese Vowels in Connected Speech

Ken'iti Kido
Professor,Research Institute of Electrical Communication,
Tohoku University
Hideki Kasuya
Research Assistant,Research Institute of Electrical
Communication, Tohoku University
Hisayoshi Suzuki
Assistant Professor, Research Institute of Electrical
Communication, Tohoku University

Introduction

The formant frequencies which are extracted from a certain part of the vowel in connected speech distribute over a wide range on F_1-F_2 plane and there are a few overlaps of the ranges of some vowels.[1)2)] The boundaries of the vowel should be separated distinctly for the discrimination of vowels in connected speech.(How to discriminate vowels in Japanese words is the object of our research.) It is necessary to take the contextual effects into account. The shifts of the formant frequencies of the vowel in /CVC/ context have already been quantitatively descrived.[3)] However,it is observed in our experiments that the formant frequencies of V_2 in /$V_1C_1V_2C_2V_3$/ context are influenced not only by the adjoining consonants(C_1 and C_2), but also by adjoining vowels($V_1$and V_3).[4)] The researches have been carried out to formulate the influences of adjoining phonemes. In this paper, some results of the acoustic analysis of the symmetrical nonsense words /hV_1CV_2CV_1/ uttered by two speakers(male adults) and the rule to discriminate the vowel V_2 are described. Where, C:/b,d,g/,V_1:/i,o/,V_2:/i,e,a,o,u/. The formant frequencies are measured from soundspectrograms. The symbols used in this paper are illustrated in Fig.1.

Analysis and the rule for discrimination

The F_{1m} vs. F_{2m} representation of the vowel V_2 in /hV_1CV_2CV_1/ context is illustrated in Fig.2. There are systematic shifts in the formant frequencies of V_2 depending on the places of articulation of the consonant C and the vowel V_1. F_{2m} of V_2 is shifted toward the value of F_2 locus of the consonant C when V_2 is a back(/a/ or /o/)vowel or a mid-vowel(/u/). That is, the value of F_{2m} of V_2 in /hV_1dV_2dV_1/ context is the highest and that in /hV_1bV_2bV_1/ context is the lowest. (F_2 locus of C_1 is influenced by V_1 and V_2.[5]) When V_2 is a front vowel (/i/ or /e/), there can not be observed the systematic shifts in $F_{2m}$of V_2 by the consonants.

The value of F_{2m} is also influenced by the vowel V_1 and its magnitude of the shift is comparable to that by the consonant. When $V_2$is a back or mid-vowel, the tendency of shifts in F_{2m} by the consonant is not influenced by the vowel V_1.

The F_{1m} of V_2 is also shifted by V_1. That is, the first formant of V_2 is shifted from its target frequency to that of the first formant of V_1. The degree of the influence of C and V_1 on the second formant frequency of V_2 can be formulated by inspecting the relation between F_{2m} and F_I. Where F_I is the average of the starting frequency and the final frquency of the second formant of V_2,and F_I should have the information about C and V_1. The value of F_{2m} has an approximately linear relation with the value of F_I,(see Fig.3)

$$F_{2m} = kF_I + R \quad . \qquad (1)$$

Where R is a constant depending on the kind of the vowel V_2, and k is a constant irrespective of the kind of V_2. Therefore,F_2 can be computed from the values of F_{2m} and F_I,

$$F_{2m} = R + R' = F_{2m} - kF_I + R' \quad . \qquad (2)$$

where R' is a constant.

F_{2m} and F_2 (normalized F_{2m}) are compared with each other in Fig.4. The fluctuations of the measured frequencies of the second formant of each vowel can be largely decreased.

Discrimination

Following the results described above, the discrimination process of the vowel in the word is illustrated in Fig.5. The utility of this discrimination process can be shown by using a linear discriminant function which is applied to the discrimination between the vowels /a/ and /o/ with 24samples analyzed in the preceding section. When F_{1m} and F_{2m} are used, the rate of the correct discrimination is 92%. On the contrary,when F_{1m} and F_I are used,the rate is 100%.

Acknowledgement

The authors wish to express their gratitude to T.Matuoka for his valuable assistance in the experiments.

References

(1) K.N.Stevens and A.s.House, J. Speech Hearing Res. 6, 111-128 (1963)
(2) N.Umeda,Nippon Telegraph and Telephone Corp. Elect. Commun. Lab. Rept. (July 1957)
(3) K.N. Stevens, A.S.House, and A.P.Paul, J.Acoust. Soc. Am. 40, 123-132 (1966)
(4) H.Kasuya et al, Paper 1-2-7, Meeting Record, Acoust. Soc. Jap. (April 1968)
(5) S.E.G.Ohman,J. Acoust. Soc. Am. 39, 151-168 (1966)

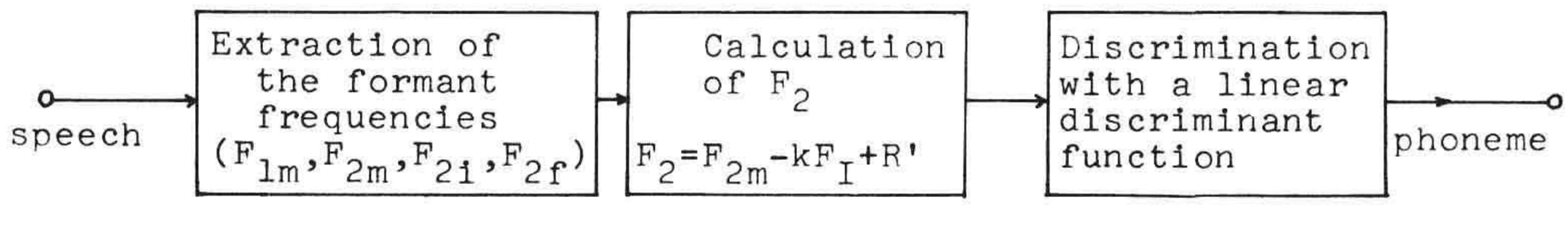


Fig.5 Block diagram of the discrimination process

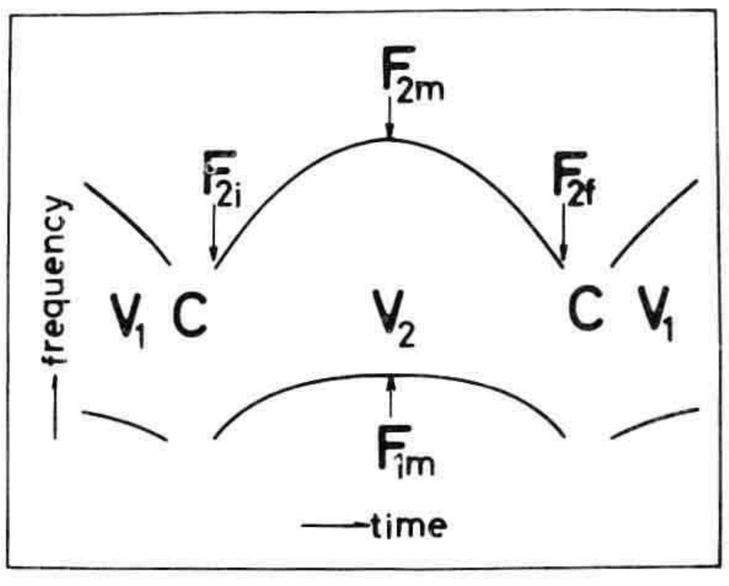


Fig.1 Symbols of the formant frequencies used in this paper.

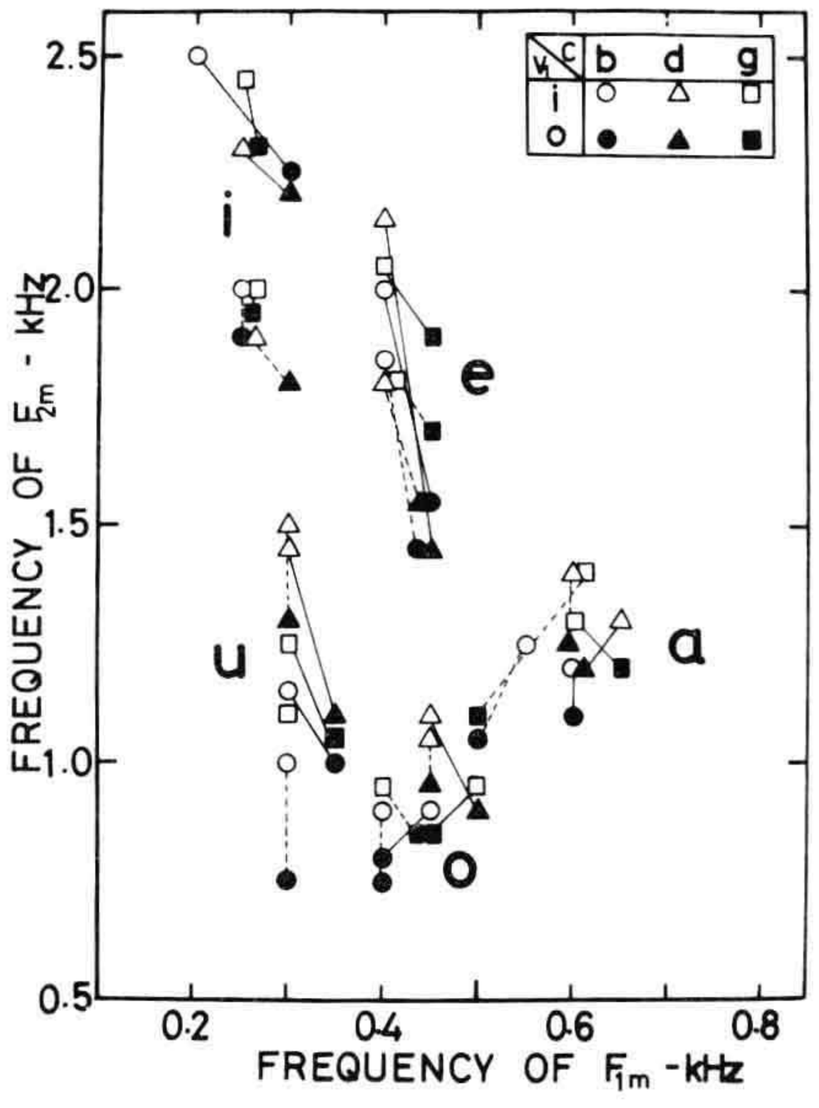


Fig.2 F_{1m} vs. F_{2m} representation of V_2 in /hV_1CV_2CV_1/ context.

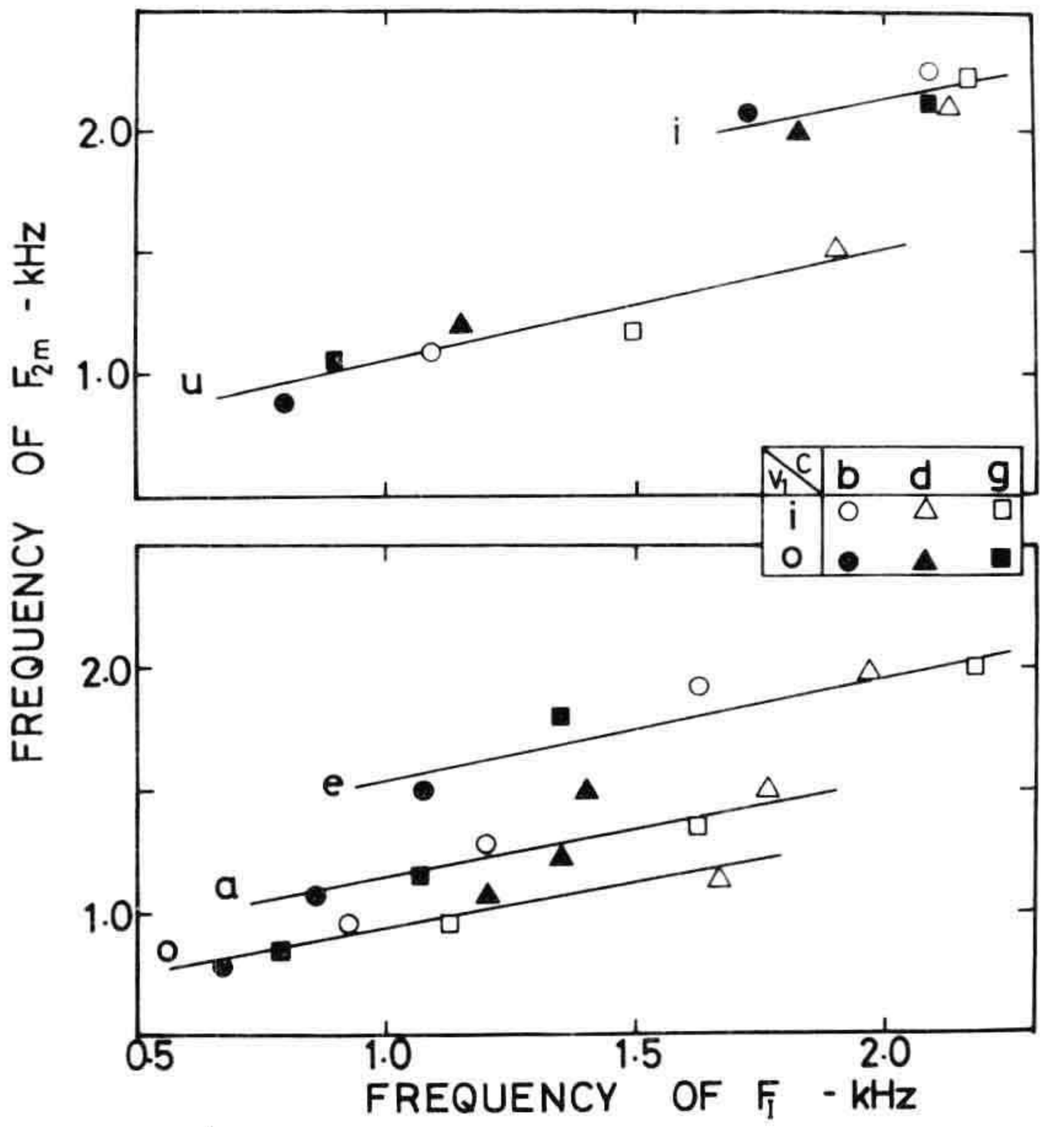


Fig.3 Relation between F_{2m} and F_I

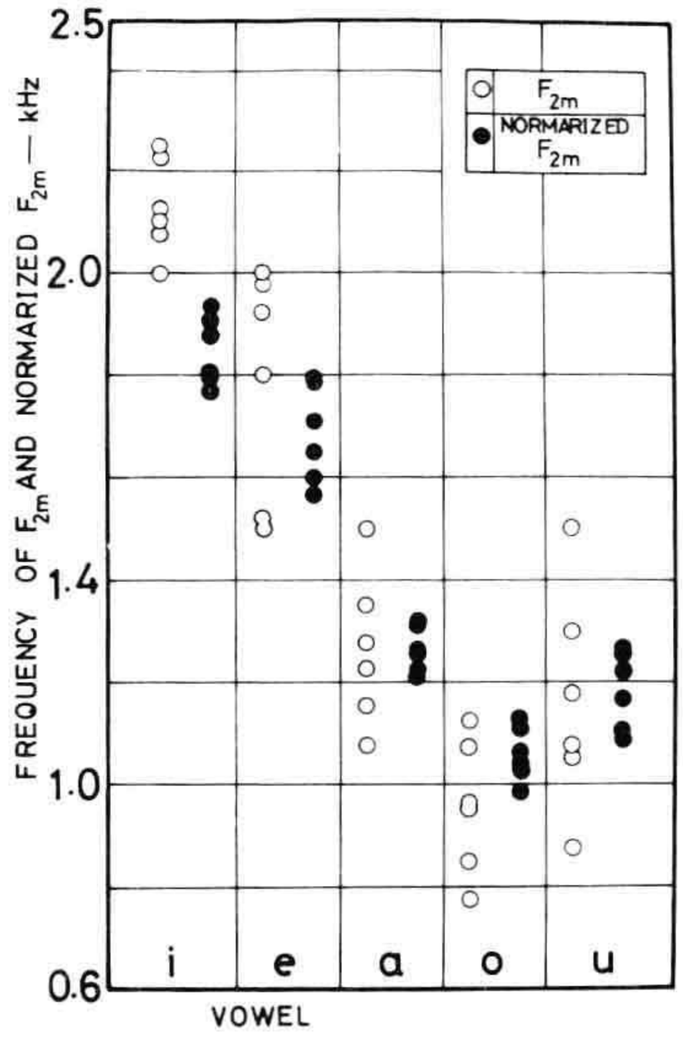


Fig.4 Comparison of F_{2m} with normalized F_{2m} (F_I)

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-4-7

A spoken digit recognition scheme and its computer simulation

Shigetaka Washizawa
B. Eng. of Nippon Columbia Co., Ltd.
Michio Okamoto
B. Eng. of Nippon Columbia Co., Ltd.
Takeo Shiga
Dr. Eng. of Nippon Columbia Co., Ltd.
Syuji Inomata
Dr. Principal Scientific Official of Electrotechnical Lab.

Introduction

A pattern-matching scheme for the recognition of male and female digit speech has been proposed. And the recognition scores have been evaluated by digital computer simulations. In this scheme, in order to normalize speech envelope patterns logarithmic compression and logarithmic time scale are used concerning diviations in level and time scale.

It has been shown that the recognition score is improved by the adjustment of weighting cofficient and by learning.

Input device for the computer-simulation

Speech samples which were tape-recorded when male and female speakers read rundam numbers in an unechonic-chamber have been utilized. The speech samples were played back from the tape-recorder. And these speech signals were recorded on recording paper by an electromagnetic oscilo-recorder. In the above process, speech signals were fed through a 3 dB/oct high-pass filter in order to stress consonants, through a variable band-pass filter in order to set an appropriate channel band-width, through a logarithmic amplifier for the suppression of level diviations, and through the other two devices to the electro-magnetic oscillograph. (cf, Fig 1)

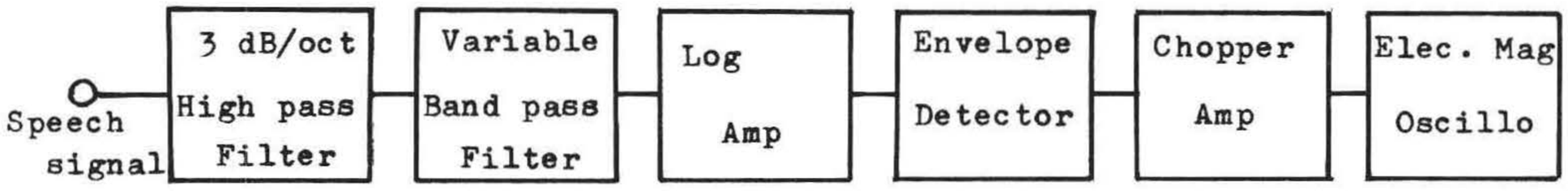


Fig 1

In the Table-1, the band-width for each channel filter, attennation factors, and time constants of low-pass filters for envelope detection are shown.

chan	Freq. range	Bandwidth	Att. factor	Time Const.
1	100 ~ 420 c/s	320 c/s	12 dB/oct	10 m sec
2	600 ~ 730	130	12	10
3	1260~1380	120	12	10
4	2000~2400	400	12	7
5	5000~5500	500	6	7
6	100~5500	5400	6	10

Table - 1

Recognition scheme

In this recognition scheme, we obtain six speech envelope patterns and sample them on twenty equally spaced-sample points on logarithmic time scale. And the recognition is performed by comparing the above samples with stored standard samples, and by finding the minimum of comparing distances.

The comparing distance is described as below,

$$D(i,i') = \sum_{j=1}^{6} A_j \sum_{k=1}^{20} |F_{i,j,k} - S_{i,j,k}| \quad \cdots(1)$$

where i = 0 ~ 9 indicates each standard speech,
i' = 0 ~ 9, each recognition-speech,
j = 1 ~ 6, channel number,
K = 1 ~ 20, each time scale,
finally Aj, each weighting coefficient.

The pre-processing in level and time scale reduces the sensitivity of the comparing distance to deviations in level and time scale. And the above

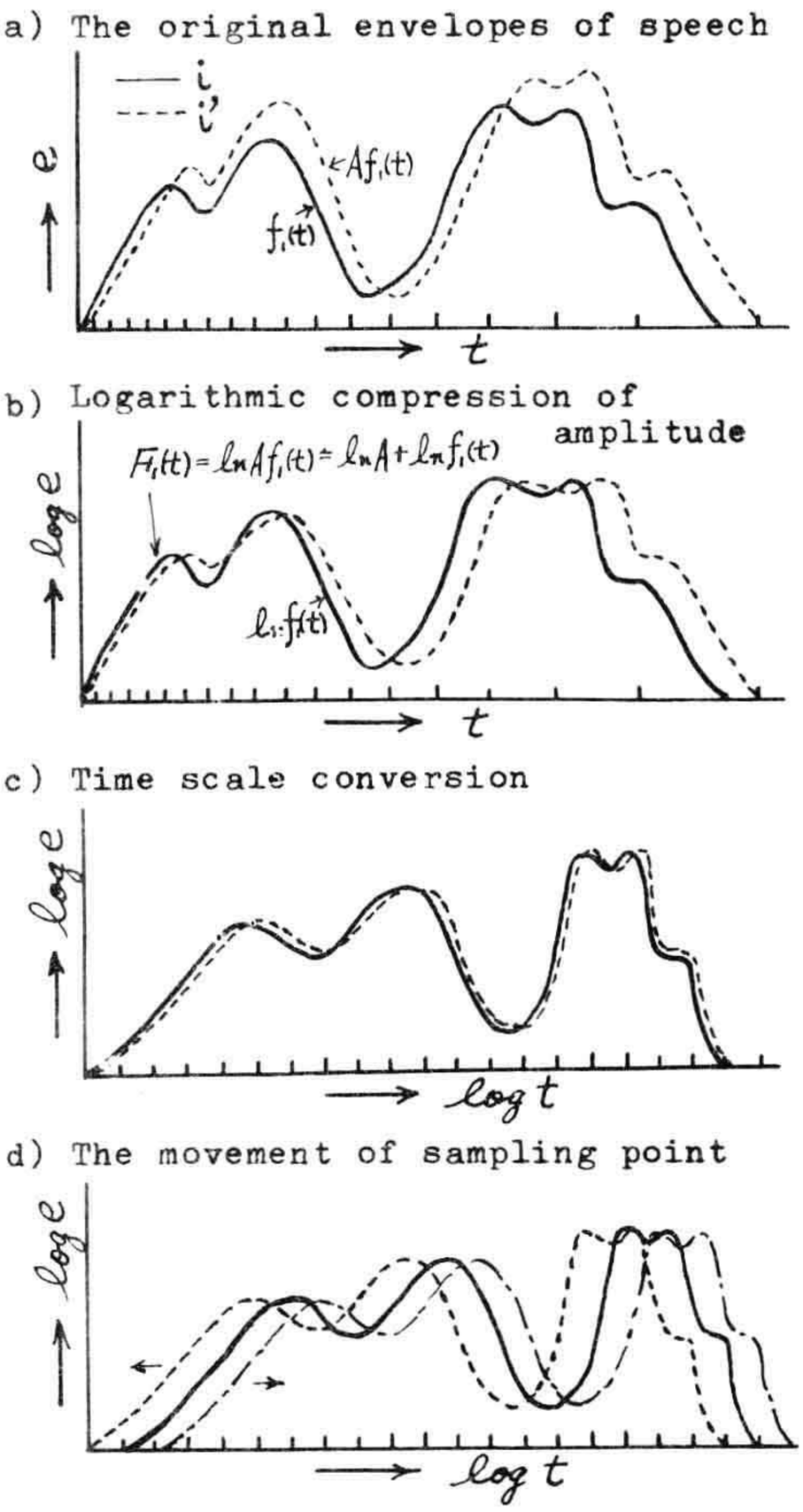


Fig - 2

situations are shown in Fig 2 (a)～(c). The time scale normalization on the other hand is performed by computing three comparing distances for the stored standard speech of three different time scales, and by picking up the minimum.

Preliminary Experiment

In order to improve recognition scores and to make recognition more reliable, the weighting coefficients in the equation (1) have been determined by Linear Programing Method by using a digital computer.

The LP method has been applied in such a way that we obtain

D (i , i') which is as large as possible when i ≠ i'

and that which is as small as possible when i = i'.

In the above requirements, a constraint such that

D (i , i') $\geqq$ d_2 when i ≠ i',

D (i , i') $\leqq$ d_1 when i = i', was used.

Some examples of the optimum of the comparing distance is shown in Fig-3.

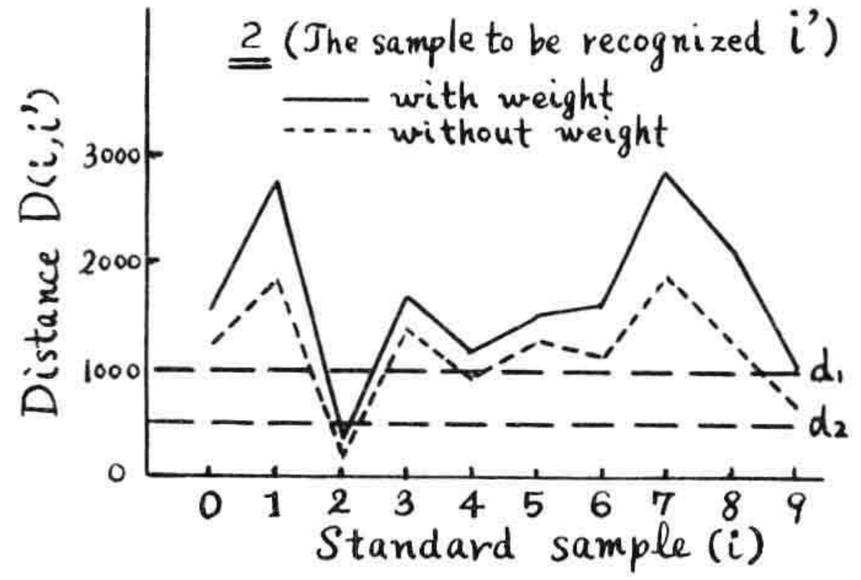


Fig 3

Computer Simulation

The proposed recognition scheme has been simulated on a digital computer and the recognition scores have been obtained. In this simulation experiment, the out put envelope patterns from the input device have been sampled and fed into a computer. And 100 digit speeches have been recognized using standard digit speech samples 0～9. In this kind of recognition scheme in which recognition is performed by picking up the standard speech sample which has the highest simularity with the input speech sample, the selection of standard speech samples is very important.

Sometimes, the standard speech samples which were originally adopted may be not optimum. And some speech samples should be discarded and new ones should be adopted in their places. In this simulation experiment, in view of the above situation, three sets of digit samples 0～9 have been adopted. And the above modification has been done at each digit at which the individual recognition score is not 100 %.

In table-2 and 4, the recognition scores of male and female digit speech without modification for 100 digit speech samples which include 10 speech

samples for each digit are shown. In table 3 and 5, the recognition scores after modification are shown. It can been seen that 94% of male and female recognition scores have been improved up to 97% for both male and female digit speech.

Conclusion

To make this scheme practical, it is necessary to store standard speech samples which are spoken by each user, and to change them as the user changes.

On the other hand it has been shown that the recognition scores are improved by normalization of speech patterns concerning amplitude and time scale. And also, it has been shown that the further improvements would be possible by introducing learning processes.

Table-2
Male, before modification (94%)

IN PUT \ OUT PUT	0	1	2	3	4	5	6	7	8	9	SCORE
0	10										10
1		8	2								8
2			10								10
3				10							10
4					9					1	9
5					1	8				1	8
6							10				10
7								10			10
8		1							9		9
9										10	10

Table-3
Male, after modification (97%)

IN PUT \ OUT PUT	0	1	2	3	4	5	6	7	8	9	SCORE
0	10										10
1		9	1								9
2			10								10
3				10							10
4					10						10
5					1	9					9
6							10				10
7								10			10
8		1							9		9
9										10	10

Table-4
Female, before modification (94%)

IN PUT \ OUT PUT	0	1	2	3	4	5	6	7	8	9	SCORE
0	10										10
1		10									10
2			10								10
3				10							10
4					9	1					10
5					3	7					9
6							10				7
7								10			10
8									1	0	10
9			2							8	8

Table-5
Male, after modification (97%)

IN PUT \ OUT PUT	0	1	2	3	4	5	6	7	8	9	SCORE
0	10										10
1		10									10
2			10								10
3				10							10
4					10						10
5						9	1				9
6							10				10
7								10			10
8									10		10
9			2							8	8

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-4-8

Some Aspects of Vowel Recognition by the Waveform

Christov Philipp
Bulgarian Language Institute,
Bulgarian Academy of Sciences

Introduction

In his fundamental work[1] S.O.Rice has shown that an effectual analysis of the random noise can be performed using the statistical distributions of some elementary waveform parameters. Later on this principle was applied to the analysis of speech[2,3,4,5]. As far as 1950, S.H.Chang[6] made the proposition for these purposes, among other parameters, to be used also the form factors of the speech wave. In this paper the form factors of five Bulgarian vowels are measured and the possible application of these acoustical data to a vowel recognition procedure is discussed.

Speech Material

The speech material consisted of five cVc nonsense syllables containing the vowels IùI, IɔI, IàI, IɛI, IìI* in position of maximum differentiation. The acoustic realization was made by the male speaker A.M., in Standard Bulgarian and recorded in a motion picture studio with professional sound recording equipment. The boundaries of the vowels were determined on the motion picture sound recordings visually as well as aurally. Segments of high-speed oscillogram were obtained by photographical enlargement of the motion picture sound recordings. The reduced 'recording speed' of the high-speed oscillogram is 32 meters per second. Each segment contained one visually determined cycle of the waveform. These segments were used for the building of waveform (WF) portrayals[8], according to the concept that a vowel can be considered as a random function, presented by a set of trials each trial being a complete cycle of the waveform.

*/ Phonetic designations according to IPA[7]

Form Factors

The form factors of the WF-portrayals of the five vowels are computed mannually. For this purpose each segment, presenting the kth waveform of the WF-portrayal, was sampled mannually with constant sampling interval $\Delta t = 100\mu sec$. From the set of samples x_{km}, obtained in this way, the mean peak $\bar{x}_{peak}$, average $x_{average}$ and root mean square x_{RMS} values of amplitude per trial were computed

$$\bar{x}^{(k)}_{peak} = \frac{x^{(k)}_{peak\ to\ peak}}{2} \; ; \quad x^{(k)}_{average} = \frac{1}{n}\sum_{m=1}^{n} \left| x_{km} \right| \; ; \quad \left(x^{(k)}_{RMS}\right)^2 = \frac{1}{n}\sum_{m=1}^{n} x^2_{km}$$

The mean values of the crest factor and the form factor per WF-portrayal are computed accordingly

$$\bar{C} = \frac{1}{k}\sum_{k=1}^{h} C_k = \frac{1}{k}\sum_{k=1}^{h} \frac{\bar{x}^{(k)}_{peak}}{x^{(k)}_{RMS}}$$

$$\bar{F} = \frac{1}{k}\sum_{k=1}^{h} F_k = \frac{1}{k}\sum_{k=1}^{h} \frac{x^{(k)}_{RMS}}{x^{(k)}_{average}}$$

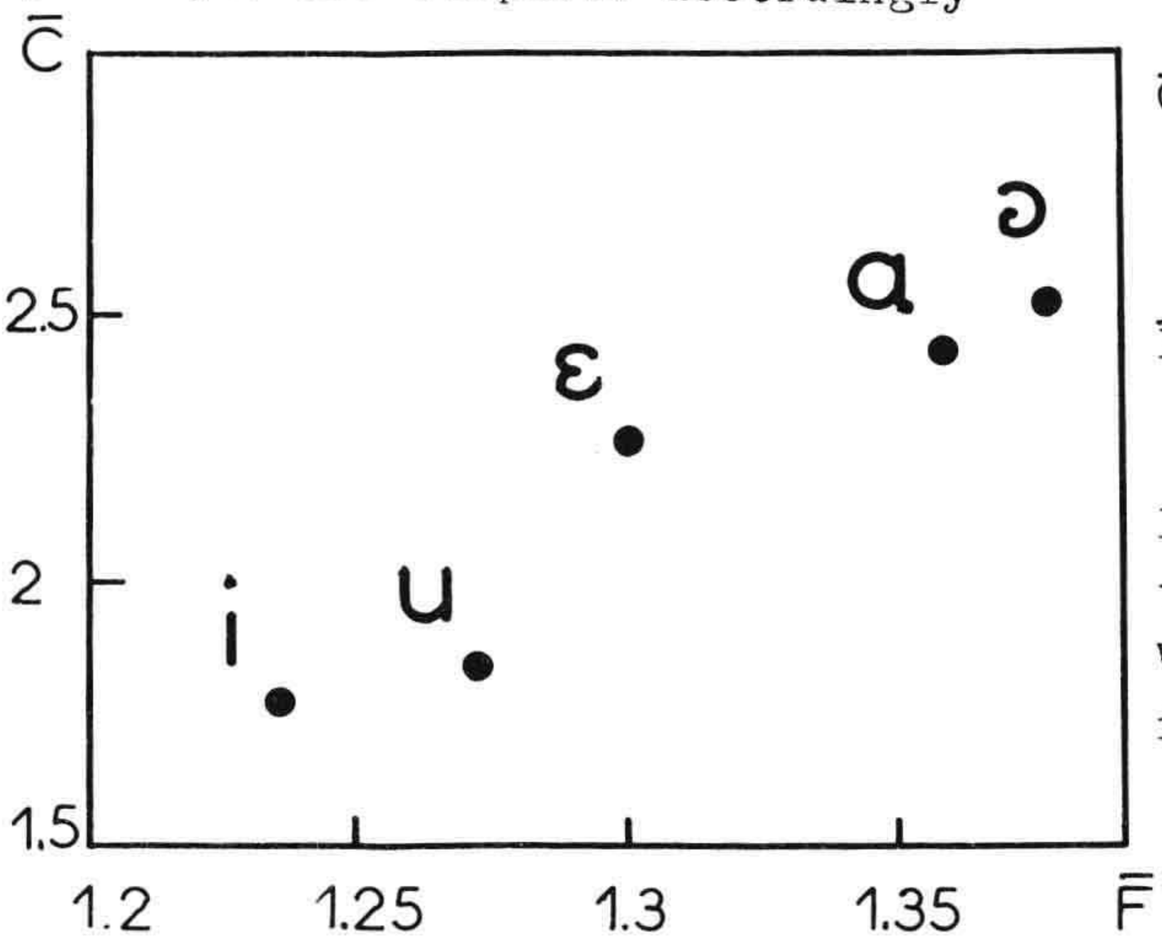


Fig.1. Acoustic representation of five Bulgarian vowels in a diagram of crest factor versus form factor

The results (Fig.1.) show that, in this example, between the crest factors and the form factors of the five vowels there exists an approximatelly linear dependence. Hence there is no reason to use a two-dimensional representation while the same information about the wave shape can be represented equivalently with the product of the two form factors

$$\bar{P} = \bar{C}.\bar{F} = \frac{\langle \bar{x}_{peak} \rangle}{\langle x_{average} \rangle}$$

which we shall call product factor of the waveform. It seems that the values of the product factor match the complexity of the waveform rather well(Ref.8., Fig.4.). Besides the new form factor $\bar{P}$ is more easily to be computed than each of the two form factors $\bar{C}$ and $\bar{F}$ taken apart.

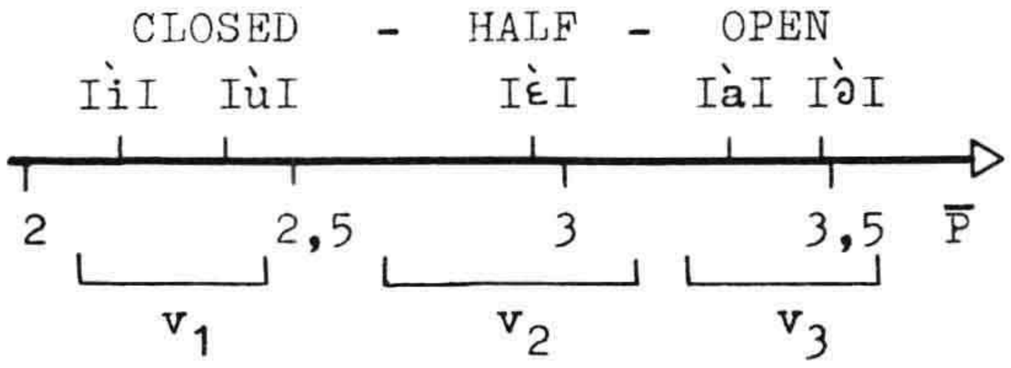


Fig.2. Acoustic representstion of five Bulgarian vowels in an unidimensional diagram based on the product of the two form factors

Curiously enough,in the measurement space of the product factor (Fig.2.) the vowels become groupped in the same manner as by the use of the articulatory criterion determined by the passage between palate and tongue(See Ref.7.,p.104). On the other hand Fig.2.,shows that, while sufficient spacings exist between the vowel classes v_1(closed), v_2(half closed/open) and v_3(open),there are to be expected difficulties by discriminating between the members of each class v_ξ .

Discussion

It is acknowledged that the results of the acoustical analysis can be highly improved by the use of linguistic information[9,10/]. In Fig.3.,is shown how much even such elementary knowledge about the language,as the frequency of occurence of vowels in speech[11/],may contribute to a vowel recognition procedure employing the product of the form factors as acoustical cue. Though this kind of linguistical statistics is not directly applicable to a short time recognition procedure,it is of use by the determining of the upper limit of the probability of false recognition of the vowels belonging to each vowel class v_ξ :

Let us suppose that the differentiation among the members $V_{\xi 1}$ and $V_{\xi 2}$ of each class v_ξ is completely impossible. Then we can lay down the rule

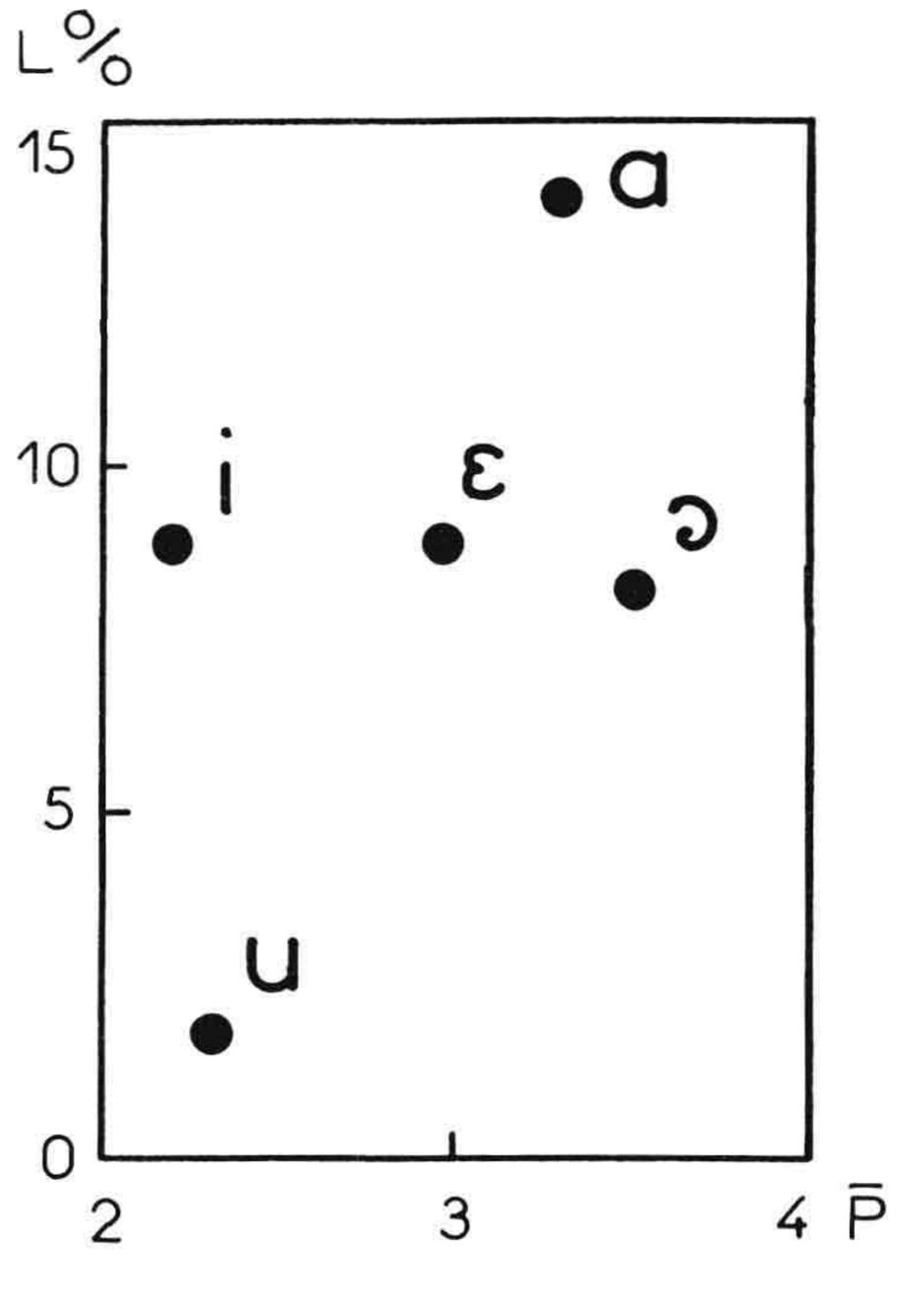


Fig.3. Representation of five Bulgarian vowels in a diagram of frequency of occurence of vowels in speech versus product of the form factors

*/ Logical symbols according to ALGOL60 [12/]

IF $(V_{\xi 1} \vee V_{\xi 2}) \subset v_{\xi}$ THEN $V_{\xi 1}$ .....(1)

where $V_{\xi 1}$ is the vowel with the highest frequency of occurence among the members of the class v_{ξ}. The probability that in this most unfavorable case the remaining members $V_{\xi 2}$ of the class v_{ξ} should be recognized false as $V_{\xi 1}$ is given in Table 1.

Articulatory classification	Closed	Half - closed/open	Open
Acoustical class v_{ξ}	v_1	v_2	v_3
Members	Ìi I, Ìu I	Ìɛ I, Ìə I*/	Ìa I, Ìɔ I
Application of rule (1)	Ìi I	Ìɛ I	Ìa I
Upper limit of probability of false recognition	17%	20%	38%

Table 1.

*/The form factors of this vowel are not computed but it is put in the class v_2 on the basis of the articulatory classification

Since the probability of false recognition, as shown in Table 1., is not very high, an error corecting procedure using higher order linguistical statistics may prove to be more effectual than the introducing of more acoustical cues.

Conclusion

A recognition procedure using the form factors of the speech wave as acoustical cues and employing statistical information about the language is supposed to be efficient by the identification of the Bulgarian vowels.

References: 1)Rice,S.O.,"Mathematical Analysis of Random Noise",Bell Syst.Tech.J.,23,282-332(1944); 2)Chang,S-H.,Pihl,G.E.,Essingmann,M.W., "Representation of Speech Sounds and Some of Their Statistical Properties",Proc.I.R.E.,39,147-153(1951); 3)Davenport,W.B.,Jr.,"An Experimental Study of Speech-Wave Probability Distributions",JASA,24,390-399(1952); 4)Sakai,T.A.,Inoue,S-I.,"New Instruments and Methods for Speech Analysis",JASA,32,441-450(1960); 5)Reddy,D.R.,"An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave",Ph. D. Thesis,Tech.Rept.NoCS49,Computer Science Department,Stanford University(1966); 6)Chang,S.H.,"Portrayal of Some Elementary Statistics of Speech Sounds",JASA,22,768-769(1950); 7)Dictionary of Linguistics, P.Owen Ltd.,London(1958),p.104; 8)Christov,P.,"A Statistical Approach to Speech Analysis by Waveform",Proc.6th ICPS,Praha(1967)(To be printed); 9)Fry,D.B.,"Theoretical Aspects of Mechanical Speech Recognition" J.Brit.I.R.E.,19,211-218(1959); 10)Tarnoszy,T.,"Can the Problem of Automatical Speech Recognition be Solved by Analysis Alone?",Rapport 5e ICA,2,371-387(1965); 11) Маринова М.,Маринов.Ас.,"Статистически изследвания на фонемите в българския книжовен език",Български език,14, 173-179/1964/; 12)Backus,J.W.,a. others,"Revised Report on the Algoritmic Language ALGOL 60", перев.англ.,из-во Мир,Москва,1965,стр.35

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-4-9

Articulatory Segmentation for Automatic Recognition of Speech

P. B. Denes and T. G. von Keller
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey, USA

Introduction

Those who try hard to produce better synthetic speech always have one consolation in their moments of dejection: automatic speech recognition is an even more intractable problem than synthesis. The great variety of schemes that have been tried to achieve automatic speech recognition all yielded more or less comparable results: good recognition could be achieved only if the words were pronounced in isolation, if the vocabulary consisted of no more than about 30 words, and if the number of speakers was less than 20 or so. New knowledge now available about the speech process encouraged us to take a new look at the problem, to review the acoustic variabilities that cause this relative lack of success, to select procedures that might prove successful, and to carry out experiments to test these new strategies. This paper describes the first of these experiments.

Articulatory Segmentation

Typical recognition schemes of the past tried to find direct correspondences between acoustic features and phonemes (or words). However, relatively little attention was given to the way in which the acoustic signal to be decoded was influenced by the special properties of its source: the human vocal mechanism. Nowadays we are much more aware of the effect of context on the physical realization of a phoneme. Some of these contextual effects are a function of linguistic factors but many others depend on articulatory influences, on "coarticulation," and on the dynamics of movement of the individual articulating organs. As a consequence, different utterances of the same phoneme may vary in ways that are difficult to identify by statistical methods but which are readily decipherable when the dynamic properties of the vocal organs are taken into consideration. For example, the spectral patterns of the same word spoken at different speeds can result in significantly different spectral patterns. Such variations may well produce greater statistical similarities between the spectra of a certain word spoken at one speed and another word spoken at a different speed, than between those of the same word spoken at different speeds. Understanding of the appropriate articulatory dynamics could help to

identify key elements and to equate two utterances of the same word whatever the speed.

As a first step towards automatic recognition, therefore, we attempted to divide the acoustic signal into segments corresponding to major articulatory events; at a later stage these articulatory segments could be integrated into phonemes or words by applying appropriate phonetic and other linguistic information. More specifically, our aim was to investigate the feasibility of identifying the boundaries of different "manners" of articulation. This would be followed by a similar study of "place" of articulation. Manner of articulation was selected as the starting point because it carries a greater functional load[1] and because determining the place of articulation would matter most at the manner of articulation boundaries.

It was inevitable, of course, that we had to use our knowledge of the articulatory phonetics of the test words for determining the necessary articulatory framework for our measurements because direct observation of the position of vocal organs was not possible. It should be emphasized, however, that we were primarily interested in finding the acoustic cues for articulatory gestures rather than for words or phonemes. At the same time, the temptation to apply our articulatory results to word recognition proved too great. Certain schemes for doing this and their results are therefore also included in this report.

Segmentation Experiments

The speech material used consisted of words spoken in isolation by a number of male and female speakers. The sound waves were recorded and then digitalized at a rate of 20 K samples/second with 12 bit accuracy. All further processing was done by computer. We attempted to identify the following categories of manner of articulation: plosive, fricative, voiced fricative, nasal consonant, vowel-like (which included all segments not otherwise classified). The merits of two different methods for articulatory segmentation were tried and compared.

1. Time domain approach. The digitalized sound wave was used directly to derive a small number of simple time functions which preserved enough information for segmentation. These functions were: (a) an amplitude function of the speech wave, (b) an amplitude function of the spectrally equalized speech wave, (c) a normalized amplitude of the twice differentiated speech wave, (d) the first zero crossing of the autocorrelation function of the lowpass filtered speech wave. Using these four functions the sound wave was divided into the following kinds of segments: (a) silence, (b) unvoiced fricatives, (c) voiced fricatives, (d) plosives, (e) vowel-like. Figure 1 shows a

	silence	fricative	voiced fricative	plosive	vowel-like
silence		12			
fricative	49	159			
voiced fricative	5	10	49	1	
plosive	50	10	1	184	2
vowel-like					572

Figure 1: Confusion matrix for segmentation experiment using time domain techniques.

confusion matrix resulting from the application of this procedure to a vocabulary of 26 words (Table 1) spoken by 9 male and 4 female speakers. 87% of the 1104 segments contained in this speech material were classified correctly. The main confusions exist between fricatives and silence, and between plosives and silence, indicating that an appreciable number of short bursts of weak plosive consonants were either missed by the speech-silence indicator or were not present.

Alpha	Hotel	Oscar	Uniform
Bravo	India	Papa	Victor
Charlie	Juliette	Quebec	Whiskey
Delta	Kilo	Romeo	X-ray
Echo	Lima	Sierra	Yankee
Foxtrot	Mike	Tango	Zulu
Golf	November		

TABLE 1

2. Frequency domain approach. A short time spectrum of the sound wave was computed. From it a speech-silence indicator, a fricative-non-fricative indicator and the lowest two formants were derived. The second formant and its amplitude served to compute a nasalization indicator. Figure 2 shows the results of a segmentation experiment.

	silence	fricative	voiced fricative	plosive	nasal	vowel-like
silence						
fricative	13	41				
voiced-fricative	2	1	18	1	5	
plosive	2	1		15		
nasal					34	2
vowel-like					3	99

Figure 2: Confusion matrix for segmentation experiment using frequency domain techniques.

These indicators were used for segmenting according to manner of articulation. Figure 2 shows the results obtained for the ten digits spoken in isolation by 5 male and 4 female speakers. The main confusions exist between fricative segments and silence, and between voiced fricatives that contain a very small amount of frication and nasals since these sounds often exhibit similar acoustic features.

Recognition Experiments

The results of each segmentation experiment were, in turn, applied to word recognition. The same features that were extracted for segmentation purposes were used to recognize the words. For the time function analysis a simple recognition logic of successive binary decisions was set up using utterances of 1 female and 4 male speakers. The recognition logic could be adjusted for error-free recognition of the 26 words. When 8 further speakers (5 male, 3 female) were tried using the same logic, only 72% of the words were recognized correctly. After readjustment of the recognition logic, 95% of the words pronounced by all 13 speakers were recognized correctly. An additional experiment with a new group of speakers indicated, however, that the performance for unknown speakers had not improved. A preliminary

recognition experiment using the frequency domain data of the ten digits spoken by 5 male and 4 female speakers yielded 85 correct and 5 incorrect decisions out of 90 trials.

Discussion

The results obtained so far enabled us to come to a number of interesting conclusions. First of all, both articulatory segmentation and word identification seemed easier by visual inspection of spectrograms than by any of the programmed algorithms. Experiments have therefore been started to investigate how persons familiar with acoustic phonetics segment and identify by visual inspection of speech spectrograms. Formants in particular appeared easier to identify by eye than by program and great emphasis is being put on finding better ways of automatic formant tracking. Visual examination of spectrograms indicates that satisfactory identification of syllable-initial R segments could be made if reliable formant tracking programs were available. Time and frequency domain segmentation schemes worked equally well, but the more sophisticated schemes to be tried soon will be easier to formulate in the frequency domain.

Remarkably high word recognition rates were obtained with our relatively simple segmentation schemes. However, the discrimination thresholds had to be changed for new speakers and the spectrograms showed clearly that recognition schemes that go far beyond improved formant tracking are required to achieve reliable recognition of even a limited vocabulary spoken by a wide variety of speakers.

Reference

[1]Denes, P. B. "On the Statistics of Spoken English," JASA, Vol. 35, No. 6, 892-904, June 1963.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-4-10

Speaker Identification by Spectrographic Voiceprints

of Voiceprint Laboratories

Kersta, Lawrence G.

President

Voiceprint Identification is a newly developed spectrographic method by which people can be identified by their voices. This method may be an important aid in helping law enforcement authorities identify criminals who use the telephone to threaten bombings of airlines and public buildings, those who make obscene telephone calls, telephoned extortion threats, and kidnapers who make ransom demands. Much like fingerprint identification, which uses the unique features in the inked impressions of people's fingerprints, Voiceprint Identification uses the unique features in the spectrographic impression of people's utterances.

The two types of Voiceprints used are shown on Figure 1. They are produced on an acoustic analysis instrument called a Sound Spectrograph. The spectrograph has the capability of translating acoustic and speech sounds to a visual portrayal of those sounds. Referring now to the Voiceprints, the one shown at the left, because it portrays the resonance bars (speech formants) of the voice, is called the Bar Voiceprint. The pattern described by these bars also determines what word is spoken. In this illustration, the word is "you". The Voiceprint also exhibits physical dimensions which include "time" along the horizontal axis, "frequency" along the vertical axis, and "loudness" as a function of the density of

the pattern. The Bar Voiceprints are used to make subjective identification of people's voices and are the type exclusively used in court room presentations.

The Voiceprint shown at the right is called a Contour Voiceprint for the obvious reason of its pattern of contours. These contours are just like those which appear on a topographical map, which instead of measuring levels of altitude, measure levels of loudness. Contour Voiceprints are used for computer automated classification of speakers and provide a filing system for identified prints.

What features of Voiceprints make it possible to identify one voice from a great many others? The Voiceprints shown in Figure 2 will help you see for yourself, because you are going to make a Voiceprint Identification. There are six Voiceprints of the word "you" which were spoken by five different male speakers. One of the speakers said "you" twice. Can you detect which Voiceprints are those of the same person?

Turning to Figure 3, we find an array of Contour Voiceprints for the same speakers, again uttering the word "you". Can you identify which Voiceprints were spoken by the same person this time? Although with a little more difficulty, you were I am sure, able to identify the correct Voiceprints by using the same principles described for the Bar Voiceprints.

Attempts to disguise the voice by whispering, muffling the voice, or holding the nose, do not alter the basic identifiable features in a person's Voiceprint. Voice mimics and ventriloquists were also unable to prevent proper identification of their voices. It is believed that each voice is uniquely different enough to make it identifiable with the same accuracy that fingerprint identification enjoys.

The grounds for the claim of individual voice uniqueness rest in the mechanism of speech. A schematic of the vocal tract, shown on Figure 4, illustrates the speech process. The parts which principally determine Voiceprint uniqueness

the vocal cavities and the articulators. The vocal cavities are resonators which, much like organ pipes, cause energy to be reinforced in specific spectrum areas, dependent upon their size. The major cavities effecting speech are the throat, nasal, and two oral cavities formed in the mouth by positioning the tongue. The contribution of the vocal cavities to voice uniqueness lies in their size and the manner in which they are coupled. The likelihood of two people having all vocal cavities the same size and coupled identically seems remote. A still greater factor in determining voice uniqueness is the way in which the articulators are manipulated during speech. The articulators include the lips, teeth, tongue, soft palate, and jaw muscles, whose controlled dynamic interplay results in intelligible speech.

The claim to voice patterns uniqueness then rests on the improbability that two speakers would have vocal cavities dimensions and articulator use patterns identical enough to confound Voiceprint Identification methods.

Identification experiments have been conducted for a relatively small experimental population (123) in which the accuracy of correct identification has been greater than 99 per cent for thousands of identification attempts. Direct application of the technique to many other voices in over 100 law enforcement applications supports the identification success achieved for the experimental population. Further statistical qualification is dependent upon the analysis of a much larger population, which as in the fingerprint identification, requires years of application.

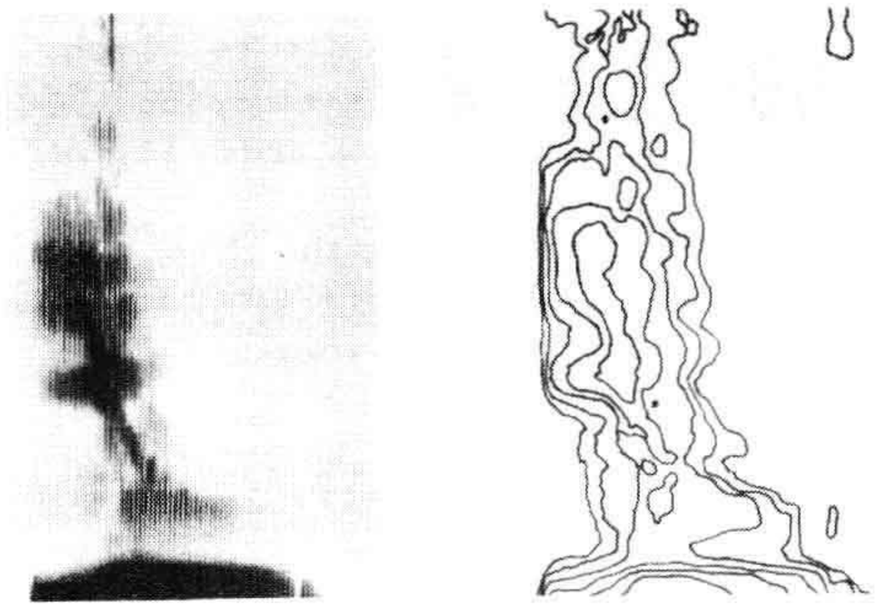


FIGURE 1
Bar & Contour Voiceprints

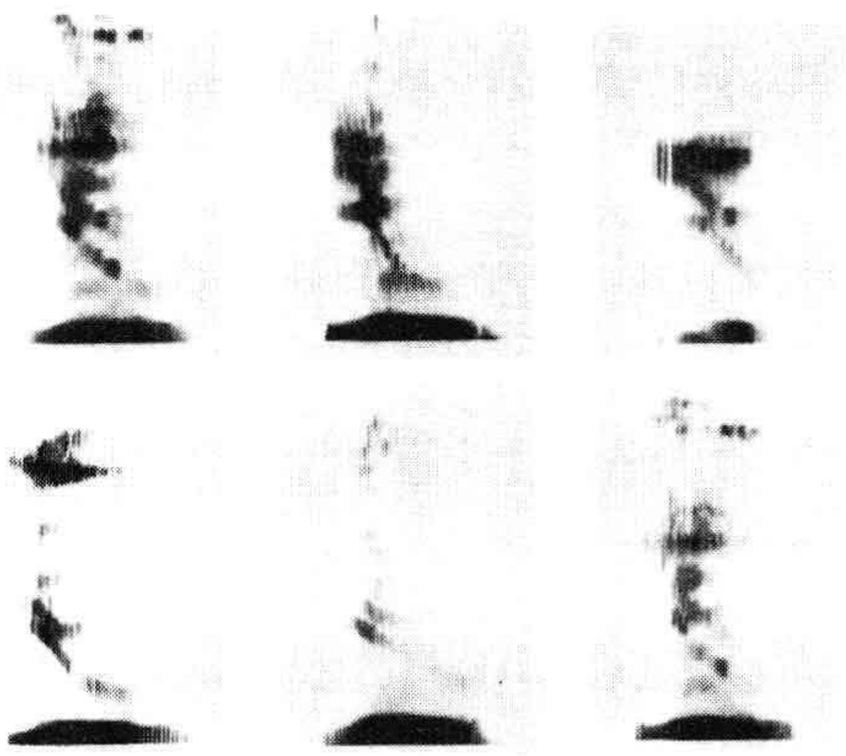


FIGURE 2
6 Bar Voiceprints

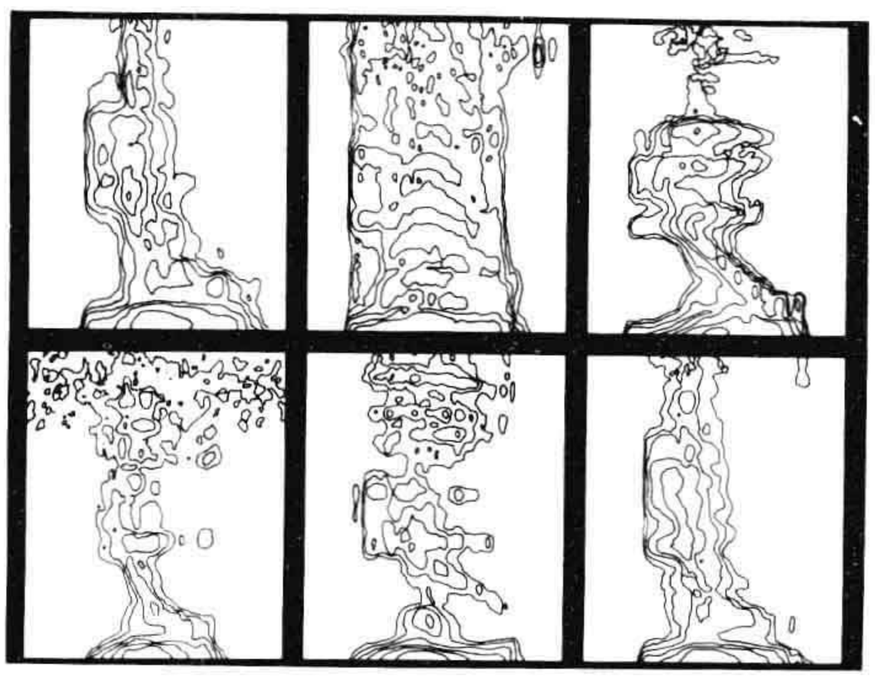


FIGURE 3
6 Contour Voiceprints

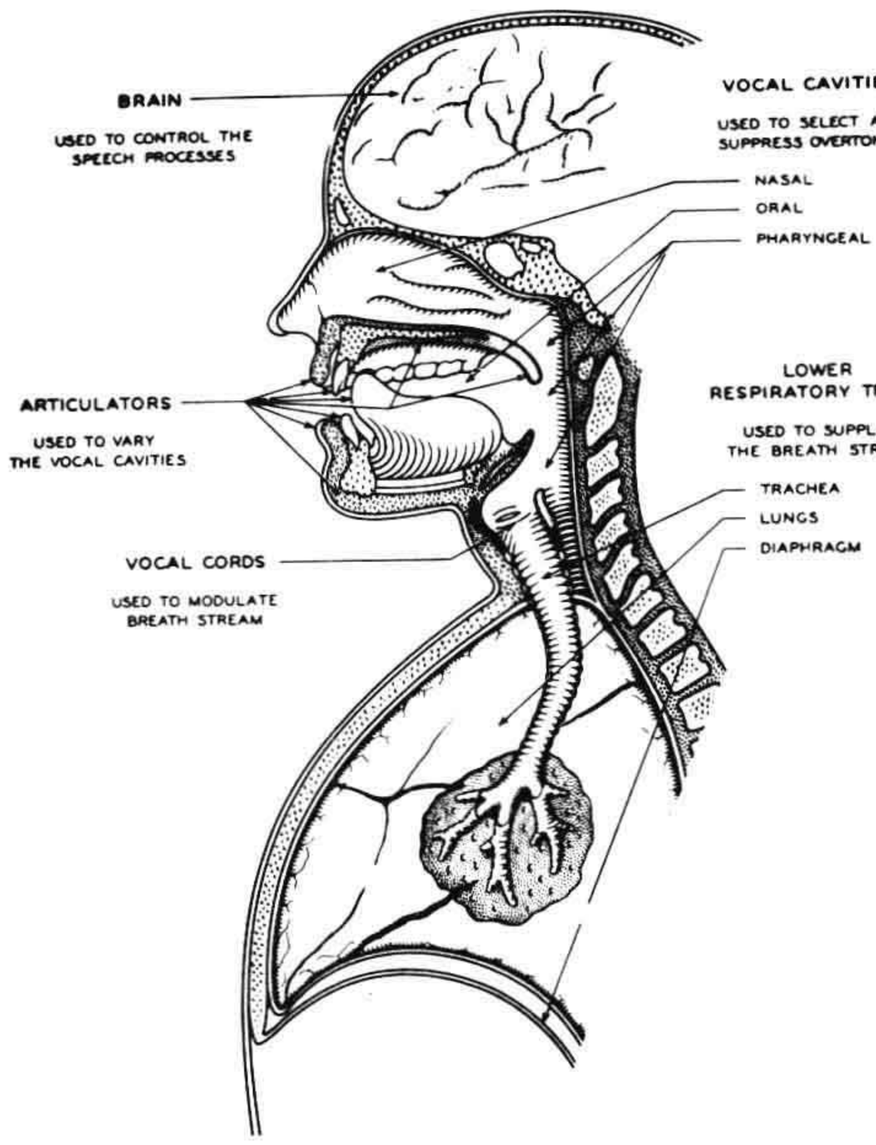


FIGURE 4
Schematic of Vocal Tract

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-1

Computer-Simulated Vocal Organs

Eiichi Matsui

Electrotechnical Laboratory

Since the acoustical model was devised in our Laboratory five years ago, we have endeavored to develop a vocal tract model using the analog computer technique. Now, we can make up a new model which depends simply on general-purpose digital computers.

The paper describes this new method of simulating vocal organs, where the transfer function of vocal tract is expressed in a form of continued fraction, and, applying the modern algorithm of the fast Fourier transform, impulse responses are calculated, then assembled pitch-synchronously into connected speech. The speech synthesizing system is also discussed briefly.

Vocal Tract Model

First we shall obtain the transfer function of vocal tract. The vocal tract is divided into m cylindrical sections of equal length ℓ, but of different cross-sectional area $A_j\ (j=1,\cdots,m)$. The acoustical behavior of the j-th cylinder can be stated in terms of a series inductance, $L_j=\rho\ell/A_j$, and two shunt capacitances $C_j/2=A_j\ell/2\rho c^2$, where ρ is density of air and c is sound velocity. Note that this approach is based on the central difference analysis of wave equations. Thus, acoustical performance of a short tube may be approximated with a π-network, shown in Figure 1, and a ladder of π-sections connected in tandem forms the non-linear transmission-line model of the vocal tract.

Let us assume that the loss factor Q of these elements increases in proportion to frequency below the critical angular frequency of ω_c, then decreases above that frequency. The assumption seems to be appropriate from several experimental data. This

suggests that the frequency parameter λ will be conveniently introduced in the following analysis, $\lambda \equiv \varepsilon + i\omega\tau/(1+i\omega\tau_H)$, where delay time of tube $\tau = \ell/c$, value of Q at 1 kHz $Q_{1kHz} = 2\pi \times 10^3 \tau/\varepsilon$, and critical frequency $\omega_c^2 \tau \tau_H = \varepsilon$ were used.

In order to obtain the transfer function of vocal tract, we shall apply Kirchhoff's law to the ladder networks mentioned above. After some calculations, transfer function $H(\omega)$, defined as the ratio of volume current at the mouth to that at the glottis, is written

$$H(\omega) = \frac{2\lambda}{i\omega\tau}\left(\frac{\ell}{\ell_{eff}} + i\omega\tau\right)\frac{2}{\Delta}\prod_{j=2}^{m}\frac{2A_j}{A_{j-1}+A_j}, \qquad (1)$$

where $\ell_{eff} = (8/3\pi)\sqrt{A_m/\pi}$ is the end correction of the vocal tract, and Δ represents the determinant

$$\begin{vmatrix} 2+\lambda^2 & -\sqrt{2}\sqrt{\frac{2A_1}{A_1+A_2}} & -\sqrt{\frac{2A_2}{A_1+A_2}}\sqrt{\frac{2A_2}{A_2+A_3}} & & & \\ -\sqrt{2}\sqrt{\frac{2A_1}{A_1+A_2}} & 2+\lambda^2 & & & & \\ & -\sqrt{\frac{2A_2}{A_1+A_2}}\sqrt{\frac{2A_2}{A_2+A_3}} & 2+\lambda^2 & & & \\ & & & 2+\lambda^2 & -\sqrt{\frac{2A_m}{A_{m-1}+A_m}}\cdot\sqrt{2} & \\ & & & -\sqrt{\frac{2A_m}{A_{m-1}+A_m}} & 2+\lambda^2+\left(\frac{\ell}{\ell_{eff}}+i\omega\tau\right)\frac{2\lambda}{i\omega\tau} & \end{vmatrix} \qquad (2)$$

This determinant appears also in eigenvalue matrix equation of the system.

Accordingly, the system function may be obtained by taking the excitation and radiation functions into consideration. Thus we have

$$S(\omega) = E(\omega)H(\omega)R(\omega). \qquad (3)$$

The above numerical analysis will easily be extended to vocal organs containing the nasal tract and also to the utterance of consonants.

Computing Sound Elements

In the preceding section, we have found the system function, i.e., the frequency spectrum of the articulatory system. The impulse response of the system is computed as the inverse Fourier transform of the system function. We assume that the ultimate unit of speech sounds be this impulse response, which will be called "sound element" hereafter. Connected speech may be assembled from a set of sound elements synchronously with pitch periods.

The computation is based on the algorithm of the fast Fourier transform, the details of which are left for references. In this paper, a method that can save half the execution time will be discussed, concentrating on the fact that a speech wave consists simply of real values.

Let Nyquist samples of temporal wave form be X_ℓ $(\ell = 0, 1, \cdots, N-1)$. The discrete

Fourier transform is defined by

$$S_r = \frac{1}{N} \sum_{\ell=0}^{N-1} X_\ell W^{-r\ell}, \quad r = 0, 1, \cdots, N-1,$$

where $W = \exp(2\pi i/N)$. If X_ℓ is real, then S_r is Hermitian symmetric, i.e., $S_r = S^*_{N-r}$.

The inverse is

$$X_\ell = \sum_{r=0}^{N-1} S_r W^{r\ell}, \quad \ell = 0, 1, \cdots, N-1.$$

If we put the even and odd-indexed samples in successive real and imaginary parts of a complex array, using Hermitian symmetry, the following Fourier transform pairs are easily shown to be given

$$(S_k + S^*_{\frac{N}{2}-k}) + i\,(S_k - S^*_{\frac{N}{2}-k}) W^k = \frac{2}{N} \sum_{j=0}^{\frac{N}{2}-1} (X_{2j} + i X_{2j+1})(W^2)^{-jk}, \quad k = 0, 1, \cdots, \frac{N}{2}-1 \qquad (4)$$

$$X_{2j} + i X_{2j+1} = \sum_{k=0}^{\frac{N}{2}-1} \left\{ (S_k + S^*_{\frac{N}{2}-k}) + i\,(S_k - S^*_{\frac{N}{2}-k})\, W^k \right\} (W^2)^{jk}, \quad j = 0, 1, \cdots, \frac{N}{2}-1. \qquad (5)$$

Thus the above relations show that the number of arithmetic operations reduces by half, because we have only $N/2$ samples to deal with.

Finally we shall estimate machine time required to compute a single discrete sound element. Most of the time for computing the system function is occupied by calculation of the determinant in the right of Equation (1). If α is average time for unit arithmetic operation which consists of recurrence formula, the time may be evaluated as $(N/2)m\alpha$. An N-point transformation requires time $(N \log_2 N)\beta$, where β is average time for complex multiplication and addition, if N is chosen to be an integral power of 2. Therefore, the time for the inverse Fourier transform in Equation (5) can be approximately given by $(N/2)(\log_2 N/2)\beta$.

Assuming that the frequency bandwidth of synthetic speech is limited to 5 kHz the duration of sound element to be computed should span about three pitch periods, say 25 msec, then we have 2^8 samples. The time required for computing a single discrete sound element is approximately 120 msec on an IBM System/360 Model 75, since $\alpha \approx \beta \approx 40\,\mu\text{sec}$. Accordingly, the time scale is about five times real-time. Thus, compared with the scatter method, our new straightforward method explained above is a little faster than the former. Moreover, we may take frequency dependent damping and radiation into consideration.

Speech Synthesizing System

Figure 2 represents schematically a speech synthesizing system based on the new

principle, i.e., the pitch-synchronous assembling of sound elements. Two files, the pronouncing dictionary and the sound element file should be prepared beforehand. First the system reads messages, written in orthographic form, and, with the aid of pronouncing dictionary, they are converted into a sequence of sound codes, which correspond roughly to phonemic symbols. Through syntactical analysis, prosodic features such as rhythm, stress, and intonation may be determined by rules. The details on this part of the system are left for the accompanying papers.

The aggregate of sound elements have to be stored on a direct access storage device. Every set of sound elements is accessed by using sound codes as a key, and assembled pitch-synchronously into connected speech. The size of inventory may be related to the quality of speech generated. A rough estimate of the number of sound elements required to synthesize English speech gives a total of approximately 2000 elements. The memory capacity needed for the sound element file is estimated at 500 kbytes. The entire program is written in the PL/I programming language on an IBM 360 computer.

The speech synthesizing system described above will develop into a powerful tool, not only for basic speech research but also for development of multiplexing voice output system which supply more than twenty lines. The synthesized speech will be demonstrated.

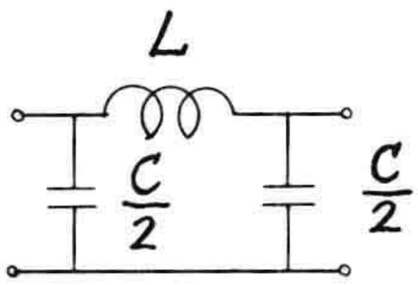


Fig 1 π-network for simulating a short acoustic tube

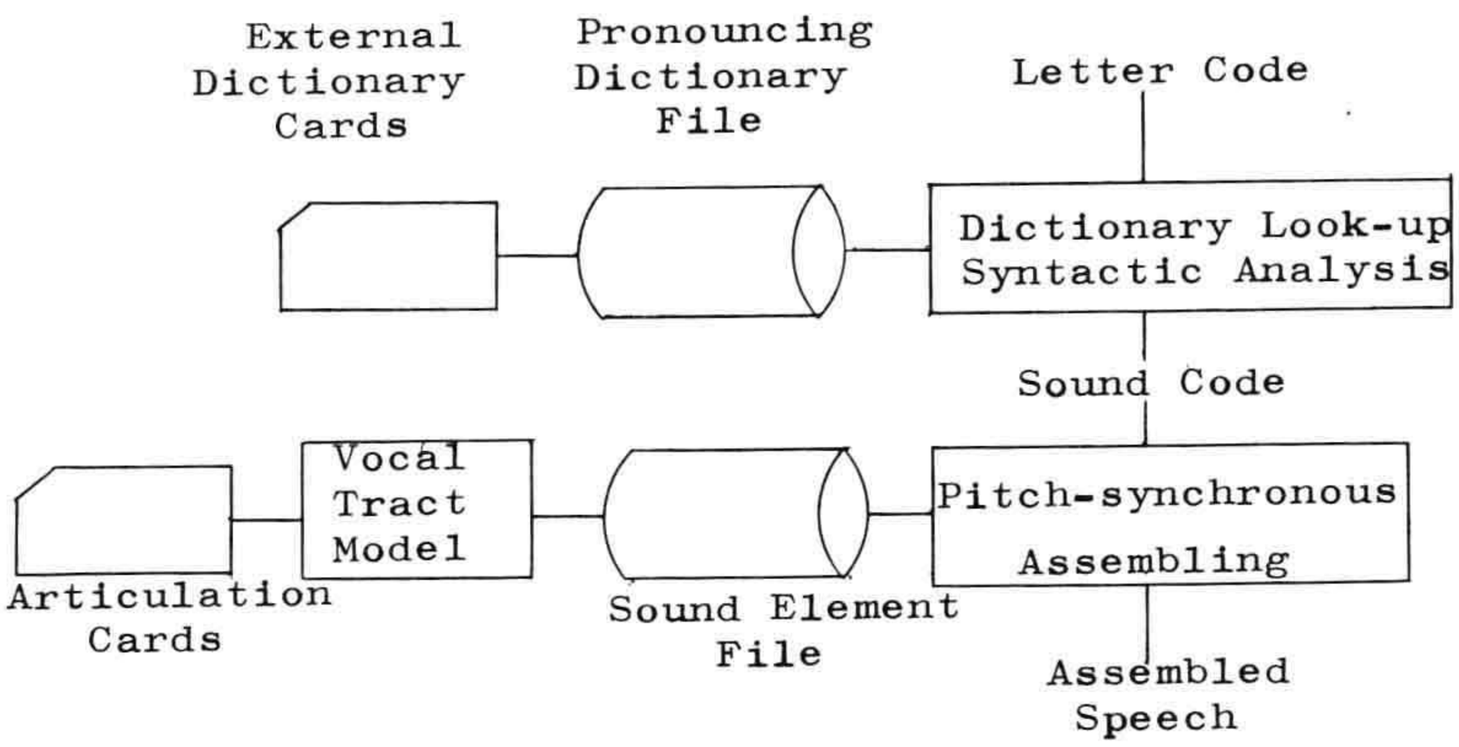


Fig 2 Speech synthesizing system based on pitch-synchronous assembling of sound elements

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-2

Use of Pronouncing Dictionary in Speech Synthesis Experiments

Ryunen Teranishi and Noriko Umeda

Electrotechnical Laboratory

Forward

The research reported here is a part of speech synthesis studies developed over several years. Overall aim of the work is to develop a system which can convert linguistic symbols into speech waveforms. Of course, many facts have been studied which concern rules covering ordinary written and spoken sentences, in order to complete such a conversion system. We had already completed a system which converts written Japanese into speech sound. However, input message for the system needed some accent marks and intonation level values which are usually omitted in ordinary written Japanese. This time we attempted to complete a system which did not need such obvious cues representing prosodic features of the sentence. Also, this time, a system was attempted which could speak a foreign language, English.

For this purpose, a "combination of dictionary look-up and synthesis-by-rule" was employed, because a written sentence generally lacks the obvious cue representing prosodic feature, and it was necessary to regenerate the cue through some technical procedure. While it is possible to minimize the dictionary, it may be impossible to omit the dictionary for reading aloud a given sentence with correct and natural prosody. The correct prosody appears as a whole effect produced by correct stress and intonation, so the point is to determine a syllable to be stressed and to find blocks composing intonation contour elements. These points were settled by use of the dictionary and grammer, an outline of the procedure is described as follows.

Dictionary Preparation

In order to use a dictionary in computer system, the desirable hardware is composed of random access memory device having a large capacity. A magnetic drum memory (IBM 2301) was used for this purpose, because all the system, except the vocal tract simulator, is realized with the IBM 360 computer. So, in our case, the drum is the most suitable device for a dictionary which contains several thousands of words. The dictionary is edited as follows. One page of the dictionary occupies one track on the drum, and it can contain 60 words when one word length is 150 bytes. If the vocabulary were 6000 words, the dictionary is about 100 pages, or 100 tracks, which is just half of the drum capacity.

Before a sentence is read into computer, it was necessary to prepare a dictionary in the system. This was arranged by the following procedure. As the first step, the vocabulary was limited to about 1500 words, tentatively, which was deemed sufficient to enable reading aloud several tales for children. Each of the extracted words is recorded on punch cards, one word per one card. The card deck exists only as an intermediary between the outer dictionary (original one) and inner dictionary (one in the machine), so the important point is to choose symbols and codes suitable for both human and machine use. A record on one card consists of letter expression as an entry, phonemic expression (with stress mark and syllabificating mark if the word is polysyllabic), and possible functions of the word.

The phonemic expression is transcribed from "Pronouncing Dictionary of American English". There are more than 26 phonemic symbols, so some of them are expressed with a combination of two letters. However, its use is only a conventional means to handling them more easily, using a typewriter. When the card deck is put into the computer, the phonemic symbol is analyzed into several allophones, using the symbol sequence information, and every allophone is expressed by an 8 bit code, the "sound code". A resultant word recorded on the inner dictionary consists of letter expression of the word as an entry, page number, phonemic expression, possible functions of the word and phonetic expression with sound codes.

Use of Dictionary

When input sentences, written orthographically with ordinary letter symbols, are presented to the computer system, they are read into the machine and divided into words. Each word is referred to the prepared dictionary, and consequently, the input literal text is converted into a series of sound codes.

The sound codes represent complete articulatory information, and a machine code for vocal tract simulator is to be derived from them. However, as they have no connection with prosodic information, some special codes have to be supplemented by the following procedure.

The series of sound codes has no break, so it must be divided into several blocks. Each of them must have adequate length and reasonable grouping, because it corresponds to a breath group, and if it is too long or too short, the speech sounds unnatural. If it ends on the halfway point of a meaning group, the speech sounds strange.

The length can be controlled arbitrarily by using a limiter in the dividing procedure. In this read aloud style, it appears that about 10 syllables are an adequate value for the limiter. The value represents maximum number of syllables contained in one block.

In order to divide a sentence into a reasonable group, the sentence is parsed into words as functional elements, refering to the dictionary. Sequence of the elements is checked from the viewpoint of syntactic pattern, and if it includes two or more sentences or clauses, the series is divided into a subseries which contains only one sentence unit. If it is still longer than the limiter value, then the subseries is again divided into a shorter subseries, according to a reasonable syntactic pattern.

Cyclic tendency, where a long and loud sound appears after a short, soft sound alternately, seems a marked feature of spoken English, so decision of stress or unstress is very important in speaking English fluently. Rules applied for deciding stress or nonstress are as follows.

It is assumed that there are only two levels i.e. stress and nonstress, so every syllable is classified into one of the two levels. Any syllable included in a "function word" is not stressed. Monosyllabic "content word" is always stressed. Polysyllablic "content word" is stressed only at a certain syllable which has been assigned a stress

mark in the pronouncing dictionary.

Here, "function word" means personal pronoun, auxiliary verb, preposition, conjunction, relative pronoun, article and so on. "Content word" means noun, verb, adjective, adverb, interrogative, indefinite pronoun, demonstrative pronoun, and so on. If a word has two or more possible functions, the real function of the word used in the environment is presumed by syntactic pattern.

All of the procedure above mentioned are programmed with PL/I language. The "Dictionary preparation" stage is expressed with about 1200 statements including "allophone process" which is composed of about 600 statements. The "Use of dictionary" stage will be expressed with about 2000 statements, most of which are already completed.

The prosodic cues in the input sentence are derived automatically using pronouncing dictionary and syntactic grammer. The result of the synthesis experiment based on the above mentioned procedure was satisfactory. More detailed conversion rules on prosodic cues into articulatory modifier, such as duration and pitch, are explained in another paper.

References

F.S. Cooper et al: Speech Synthesis by Rules. in "Proceedings of the Speech Communication Seminar, Stockholm 1962, Vol. 2" Speech Trans. Lab. R.I.T. Stockholm, Sweden (1963)

J.S. Kenyon and T.A. Knott: Pronouncing Dictionary of American English Merriam Co. Springfield, Mass., U.S.A. (1953)

C.C. Fries: Teaching and Learning English as a Foreign Language. The Univ. of Michigan Press, Ann Arbor, Mich. (1945)

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-3

Synthesis of Fairy Tales Using an Analog Vocal Tract

Eiichi Matsui, Torazo Suzuki, Noriko Umeda and Hiroshi Omura

Electrotechnical Laboratory

Introduction

This paper describes a system of synthesizing English stories, especially fairy tales, in which a vocal tract simulator controlled by a computer program is used. Letter symbols, orthographically written English sentences, are converted into intermediate symbols, which are led from reference to a pronouncing dictionary and syntactic analysis, as reported in the preceding paper. The intermediate symbols, which are the input to our system, are converted into machine codes, using articulatory and prosodic rules. The vocal tract simulater is operated with the machine codes, control data every five mili second, and synthesized speech is obtained.

System of Speech Synthesis

Intermediate symbols are used as an input to the system, led from the pronouncing dictionary and syntactic analysis. They consist of phonemic symbols with some special marks. As is shown on the figure in the last page, phonemic symbols are converted to sound codes through allophonic process. For pronouncing phonemes, their vocal tract postures are influenced to a large extent by their environment, especially upon their following vowel configurations, so an allophonic process is indispensable for speech synthesis on an analog vocal tract.

Sound codes are considered to represent target vocal tract configurations of allophones from which are drawn all articulatory data concerning sound sources and vocal tract. Application of articulatory rules to articulatory data results in

producing time charts. These charts are the basic source of instanteneous control data every five mili second.

Special marks represent prosodic rules of rhythm, stress and intonation which appear physically as duration of sound codes, pitch frequencies and intensity. The time charts produced on an articulatory level are modified by these special marks according to prosodic rules, and pause duration between blocks are decided, too. Then the machine codes, that is, the control data for operating the vocal tract simulator are computed and are written out on magnetic tape.

In our system of synthesizing speech, the control part is processed on an IBM 360 computer which is not connected on-line with our analog vocal tract. So the magnetic tape on which control data are recorded is reproduced on a magnetic tape handler to control our simulator. Vocal tract simulator consists of analog computers simulating vocal tract in real time with source excitation parts.

Sound Codes and Articulatory Rules

When a phonemic symbol is read in, an allophonic process is made. The code thus processed is called a sound code. Sound codes are decided by the following principle: they represent target vocal tract configurations, so, if a phoneme is required to have several different vocal tract shapes as allophones, according to the following sounds, sound codes corresponding to the phoneme are set as much as the difference of environments. For vowels, a phoneme corresponds to a sound code. A consonant phoneme is assumed to have three codes in general. That is, when it is followed by front vowels, it takes a configuration in which the tongue is set forward and when it is followed by back vowels, it takes a backward configuration. Neutral configuration is taken in other cases. Now eighty three sound codes are prepared in our programming system. This is the minimal number for synthesizing English speech. It is desirable to prepare many more codes (configurations) to a phoneme so as to obtain natural speech sounds with high quality. Because of simplicity of the rules and from a practical viewpoint, eighty three is considered to be a satisfactorily efficient number.

Sound codes thus decided function to draw out articulatory rules and data related to them. First, as articulatory data, target vocal tract configuration of each sound code and its standard duration time are decided according to allophonic environment, and

the stationary values of the sound sources which are considered to be proper attributes, are described. Then, these data are constructed into time series of configuration, buzz, pitch and hiss source, applying articulatory rules. Articulatory rules are those governing temporal movements to the next codes by means of linear interporation on configuration, buzz intensity and frequency and hiss intensity. Vocal tract configurations are obtained, using various means such as X-ray photographs on vocal tract, sound spectrograms and listening.

Special Symbols and Prosodic Rules

Special symbols act as prosodic modifiers of the standard time charts made by articulatory data. They are stress marks (single quotation and asterisk), separators between syllables (hyphen), words (break) and blocks or in other word breath groups (blank, comma, and full stops), and a mark indicating rising intonation. Prosodic rules such as modification of sound duration, decision of pause duration and making rising and falling pitch patterns are referred to using these special symbols.

On an assumption that stress in spoken English appears most markedly in syllable length, sound duration times are given in a table where three values are prepared for a sound code. Standard value is chosen when it appears in a syllable with single quotation mark, a shorter one is used in a syllable without any mark and a longer one with an asterisk. Time charts which are settled from articulatory rules and data are composed of stationary part and gliding parts by turn. A sound code contains a stationary part followed by a gliding part. As gliding parts are assumed to occupy 50 mili seconds in almost all cases, modification of duration time is carried out on the stationary parts.

Pauses are taken between blocks. Their values are decided according to whether the separators are blanks, commas or full stops.

Sentence pitch patterns are considered to be analyzed in several levels. Phenomena on the articulatory level, for instance frequency overshooting seen at plosive onset, are treated in the articulatory rules. Here, as one of the prosodic features, pitch patterns have to be considered on two levels; one is that within a block and the other is that of a sentence. All blocks have falling pitch pattern which comes from the spontaneous pattern of respiration. Besides, stressed words in a block act as keys for

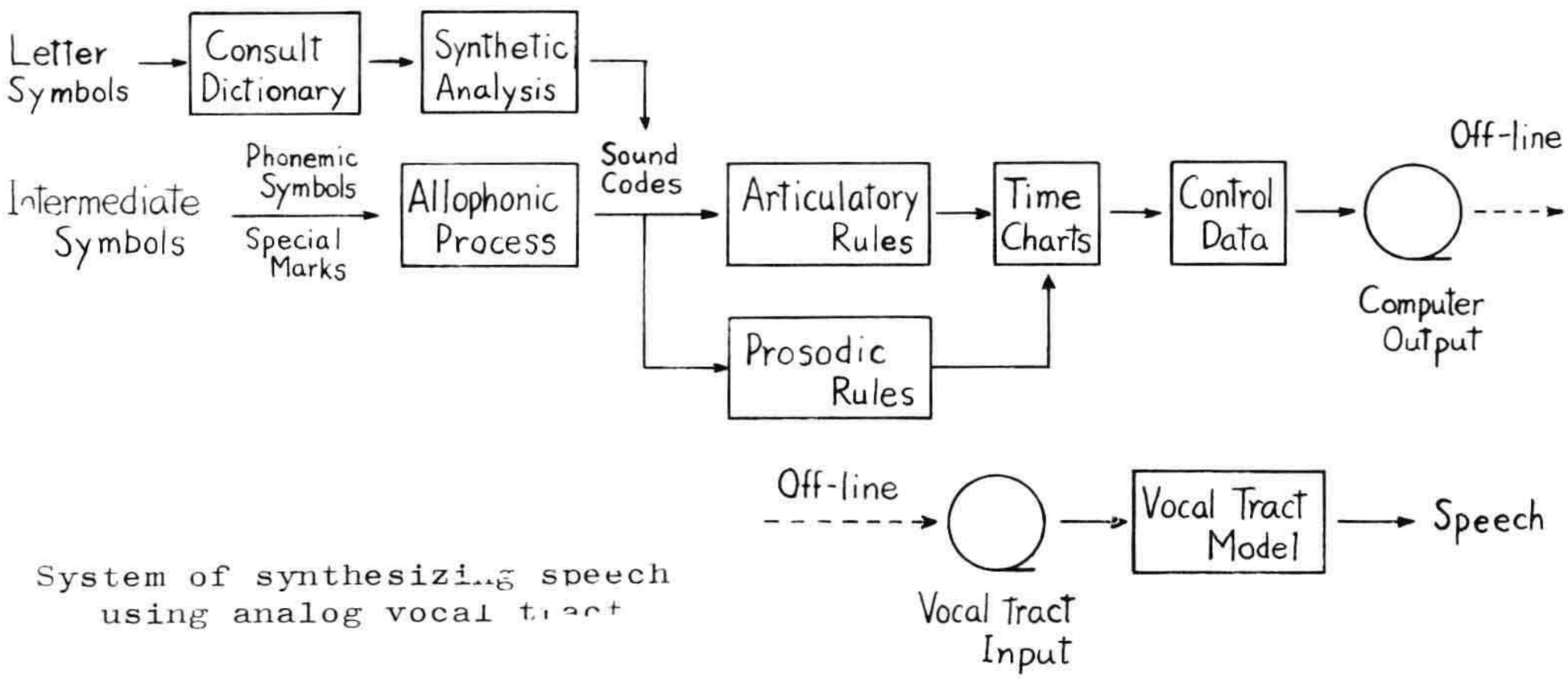
making rising and falling intonation superimposed on the falling pattern mentioned above, forming a syntactic unit. As a sentence pitch pattern, frequency levels at startpoints and end of blocks are decided according to whether they appear at the head of the sentences, at the tail or in the middle. When the mark of rising intonation is used at the end of a block, the pitch of the last syllable is raised.

Only the intensity patterns of articulatory level are considered, and prosodic level patterns are not controlled at all.

Prosodic rules are superimposed on articulatory level, and the time charts which are led from articulatory rules are modified.

Sequence of intermediate codes are thus converted to a sequence of machine codes every five mili second, which operate the vocal tract simulator, and synthetic speech is obtained. In our synthesizing system, the IBM 360 computer is used. The program is written in PL/I language.

Our present purpose is to synthesize speech of plain sentences, like fairy tales with simple and primitive rules at a moderate speed. As an example, "Sleeping Beauty" from Grimm's Fairy Tales will be demonstrated.



System of synthesizing speech using analog vocal tract

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-4

A Model of Word and Sentence Intonation

Sven E.G. Öhman

D.Sc., Department of Speech Communication,

Royal Institute of Technology, Stockholm, Sweden

The basic phonetic difference between the Swedish utterances "där ser du tȍmten" = "there you see Santa Claus" and "där ser du tómten" = "there you see the site" consists in the tonal patterns of the words "tȍmten" and "tómten".

These "tones", called the grave and the acute accents, respectively, are like phonemes in that they serve to distinguish between words. Sentence intonation, on the other hand, expresses functions such as "question", "statement", "continuation" and so forth. This type of intonation extends over many words and is not directly tied to individual word meanings.

A third factor entering into speech melody is stress. The pitch of the speaking tone rises momentarily during stressed syllables. The distribution of stress within the utterance depends in great part on syntactic structure. Sometimes the pitch modulations due to stress are also labelled "sentence intonation".

It has often been observed that word intonation is superimposed on the sentence intonation. The last word of a Swedish question, for instance, can carry any tonal accent, and, vice versa, a given tonal accent can fall anywhere on the sentence intonation. To investigate the nature of this sort of superposition a model was constructed as follows.

A fundamental frequency (pitch) signal $f_o(t)$ is synthesized by a mechanism which accepts three types of input: (1) the time varying "vocal cord tension"g(t),

which is the sum of two components, $g_s(t)$ and $g_w(t)$, where $g_s(t)$ represents the sentence intonation and $g_w(t)$ the word intonation, (2) an acoustic interaction signal arising from the secondary effects on f_o caused by fluctuations in sub- and supra-glottal pressure, and (3) an articulatory interaction signal deriving from the non-phonatory movements of the hyo-thyroid lever system. These movements are due to certain articulatory gestures of the tongue.

The two signals $g_s(t)$ and $g_w(t)$ are the outputs of two second order linear filters that may have different properties. These filters are assumed to summarize the essential dynamic characteristics of the mechanical and peripheral-neural components of the laryngeal system.

The inputs to the sentence and the word intonation filters are built up from step functions with different amplitudes and onset times. These steps represent the discrete higher-level neural events that correspond to the linguistic intonation elements (levels, tones, etc.). In particular, it is assumed that only a finite "library" of step amplitude and timing configurations may be drawn upon in the construction of the f_o-contour of an utterance. Empirical investigation involving systematic matching of the model to data was relied on to decide what these configurations are like in Swedish. The matching was done by means of a digital computer.

It was found that the essential features of simple intonation patterns such as those of [de vɑ: mó:nen ja sɑ:], [de vɑ: mò:nen ja sɑ:], etc., can be simulated by means of a single positive step as input to the sentence intonation filter together with a negative pulse as input to the word intonation filter.

The positive sentence intonation step must start at the very beginning of the stressed syllable. The negative word intonation pulse should be introduced early or late in the syllable, depending on whether the accent is acute or grave. The depth, duration and latency of the pulse varies for the two accents from dialect to dialect, but the basic configuration of a negative word intonation pulse (early in the syllable for one of the accents and late for the other one) superimposed on a positive sentence intonation step appears to capture the tonal essentials of all

Scandinavian dialects (including Danish), that preserve the accentual contrast.

In view of the negative pulse character of the word intonation command in the model, and in view of the phonetic realization of the acute accent in Danish it is suggested that the grave as well as the acute accents should be understood, phonetically, as a kind of more or less lax glottal stops in all dialects. Such glottal stops should probably be attributed to the valve function of the larynx musculature while sentence intonation probably is controlled by partly separate "singing muscles". An electromyographic pilot study [1] gave some support for this idea, but the question must be further investigated.

A reexamination of E. Meyer's[2] in terms of the model proposed above indicates certain empirical relationships between the parameter values for the acute and grave accent commands of the various dialects. In particular, although almost any latency seems to be possible for e.g. the acute accent command - in the sense that any given latency for this accent can be exemplified by at least some dialect - it appears nevertheless to be the case that the <u>difference</u> in latency for the two accent commands in any given dialect is close to a certain favored value.

In going from dialect to dialect in a certain sequence, the pitch inflection patterns associated with the two accents can be seen to shift gradually and pari passu toward the end of the stressed syllable. As the later one of the two patterns approaches the end of the syllable it becomes weaker in amplitude until it eventually vanishes. However, the pattern reappears at the beginning of the syllable in the dialect of the sequence, so that here the time order of the two accents is the reverse of that typical of the preceding dialects.

The gradual shifting of the patterns is found to continue as before, however, also from this point on in the sequence until the time order of the two patterns is switched around once more. At this point the original configuration appears anew. The whole "process" accordingly has a cyclic character. (Of course, this cycle could have been traced out in the direction opposite to the one chosen in the above description.)

The facts related here are probably due to mutual influence among the dialects. It remains to be seen to what extent they also reflect intrinsic limitations on tonal phenomena.

References:

(1) S. Öhman, B. Mårtensson, R. Leanderson, and A. Persson: "Crico-Thyroid and Vocalis Muscle Activity in the Production of Swedish Tonal Accents: A Pilot Study", STL-QPSR 2-3/1967, pp. 55-57.

(2) E. Meyer: Die Intonation im Schwedischen, Teil I (Stockholm 1937) and Die Intonation im Schwedischen, Teil II (Stockholm 1954).

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-5

Generalization of Control Rule of Buzz Source to Improve the Naturalness of Synthetic Speech

Takeshi Nakayama, Akira Ichikawa and Tanetoshi Miura

Central Research Laboratory, Hitachi, Ltd.

Introduction

A control rule of buzz source to improve the naturalness of synthetic speech was proposed at the 74th ASA meeting[1]. Further study of this rule for longer sentences has been carried out by both analysis and synthesis.

Basic assumptions

In the analysis, sequential variations of intra-mora mean pitch frequencies, source intensities and mora durations of the speech uttered by an actress were examined on the following assumption: The sequential pattern of each parameter can be expressed as the product of four independent factors H, S, A, and the basic parameter.

For example, for pitch pattern,

$$P_{iklj} = H_{(p)ij} * S_{(p)kj} * A_{(p)lj} * P_{oj} \quad \text{.....(1)}$$

where, P_{iklj} is the intra-mora mean pitch of jth mora spoken by a person (i), reading a sentence (k), in the expression (1). $H_{(p)ij}$ is the personal factor of speaker (i); $S_{(p)kj}$, the sentence factor of sentence (k); $A_{(p)lj}$, the expressional factor of expression (1); and P_{oj}, the basic intra-mora pitch of jth mora. The same assumption as in equation (1) is applicable to both the intra-mora mean source intensity, i_{iklj}, and the mora duration, t_{iklj}.

Determination of basic parameter functions

The basic intra-mora mean pitch P_{oj}, (j=1,2,...n) are determined by averaging intra-mora mean pitches obtained from several speech samples of the same number of moras of the same speaker. Figure 1 shows the observed intra-mora mean pitch of ten different speech samples. These samples were uttered by an actress and have different sentence lengths. However, as clearly seen in Figure 1, they show regular shapes common to all sentences as a function of mora. These shapes are found to be approximated by the following equations.

$$P_{oj} = P_o + \sum_{k=1}^{N} P_k \quad \text{.....(2)}$$

Where,

$$P_o = 150 + 75\ e^{-0.22j} \quad \text{.....(3)}$$

$$P_1=P_2=...=P_k=.....P_n=1200\ e^{-0.82(j-m+1)}-1866\ e^{-1.29(j-m+1)} \quad \text{.....(4)}$$

k is the numerical order of segment in the sentence, and m is the numerical order of the first mora of the segment.

Figure 2 shows the average intra-mora mean pitch of the first seven moras of ten sentences (dot), and the calculated pitch by equations (2), (3) and (4) (solid line). Figure 3 also shows the comparison between observed and calculated values for the twelve-mora-sentence. The arrow on the abscissa indicates the first mora of each segment.

Generalization of expressional factors

In equation (1), $S_{(p)kj}$ is related to the accent pattern. Therefore, its behavior is relatively known, and has already been adopted in our synthesis rule. From the practical point of view, $H_{(p)ij}$ is not of primary interest, because it will suffice to synthesize representative male, female, and if possible, child voices. On the other hand, very little is known about $A_{(p)1j}$, which is important for efficient transmission of language information by synthetic speech.

The expressional factor $A_{(p)1j}$ is given by the following equation.

$$A_{(p)1j} = (\bar{P}_{1j} / \bar{P}_{oj}) * P_j^{(1)} \quad \text{.....(5)}$$

Where, P_{lj} designates P_{iklj} for the same speaker and sentence but different expressions. P_{oj} denotes the reference value (expression as a statement). $\bar{P}_{lj}$ is the mean value of P_{lj} for all moras of the sentence. $P_j^{(1)}$ is defined by the following equation.

$$P_j^{(1)} = \frac{P_{lj} / \bar{P}_{lj}}{P_{oj} / \bar{P}_{oj}} \quad \text{.....(6)}$$

Values of both terms on the right-hand side of equation (5) for longer sentences are compared with those of eight-mora sentences. As shown in Figure 4 and Table 1, $A_{(p)lj}$ for longer sentences is approximately equal to that for eight-mora sentences.

Concluding remarks

From these results, it has become evident that the rule based on the analysis of eight-mora sentences is applicable for longer sentences.

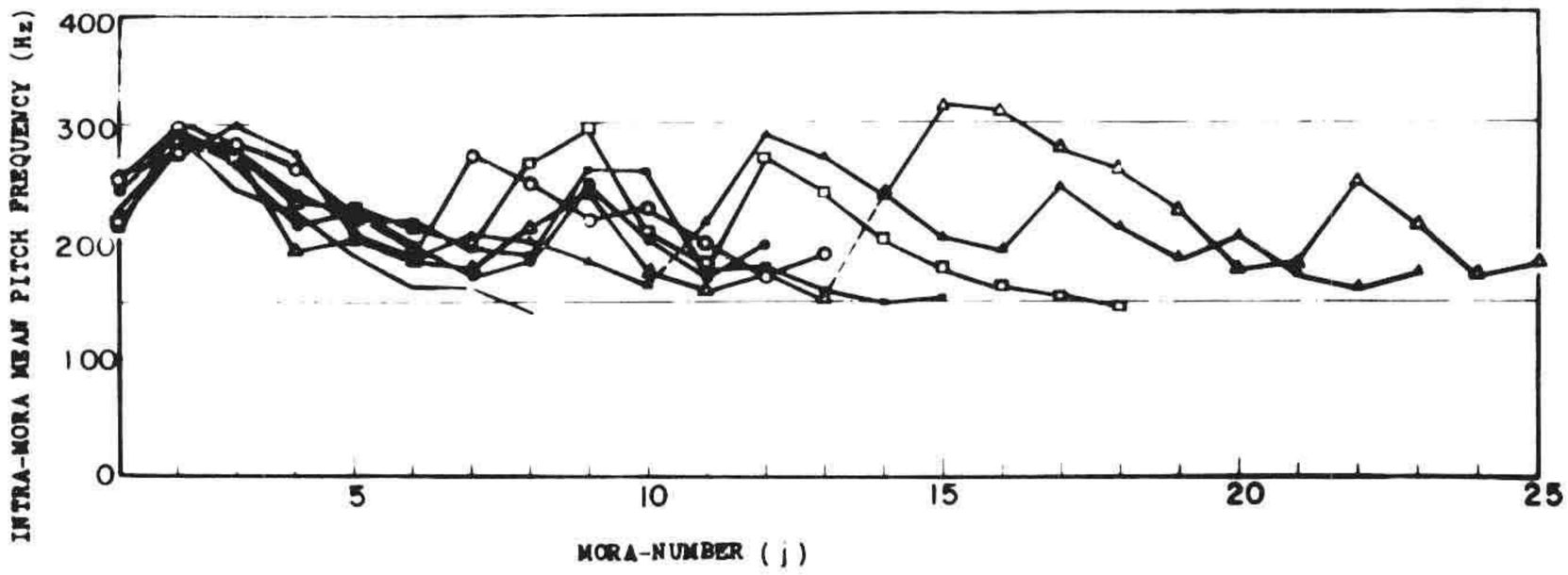


Fig. 1 Observed intra-mora mean pitch as a function of mora.

— Mean value for four eight-mora sentences.
yūyakezorani hoshigadeta
ōnamiga isoni uchiyoseta
yaoyano oiga aoí ieo uru
urano hatakede mugino aoi mega nobiru
taifūno eikyode ōnamiga isoni uchiyoset
nagaiaidano amega agari yūyakezorani hoshigadeta
---- pause

Table 1. $\bar{P}_{ij} / \bar{P}_{oj}$ for various expressions

short : aoi-ieo-uru
long : yaoyano-oiga aoi-ieo-uru

sentence expression	short	long
statement	1.027	0.995
yes-no question	1.176	1.094
emphasized statement (1)	1.036	1.055
emphasized statement (2)	1.118	1.139
emphasized statement (3)	1.229	1.104

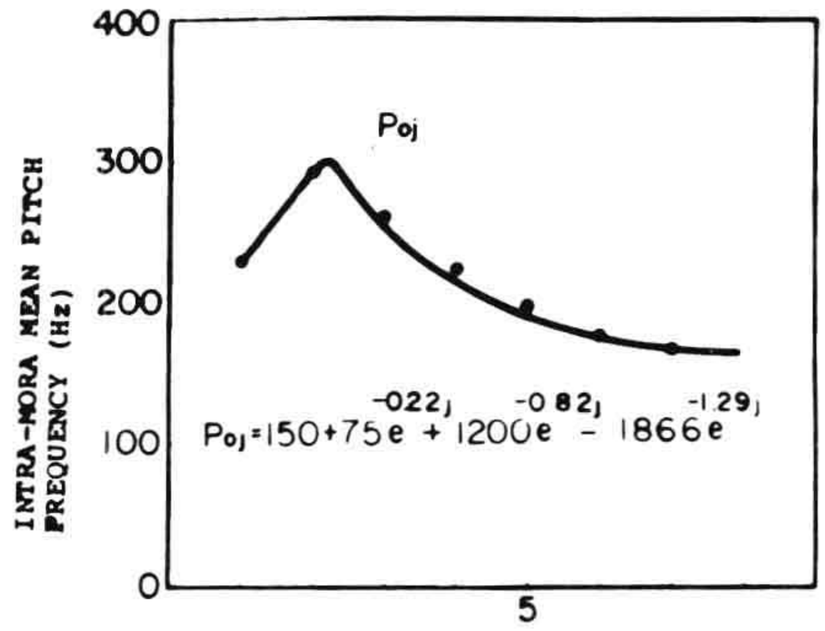


Fig. 2. Comparison of observed and estimated intra-mora mean pitch frequencies. Dots show observed values and solid line estimated values.

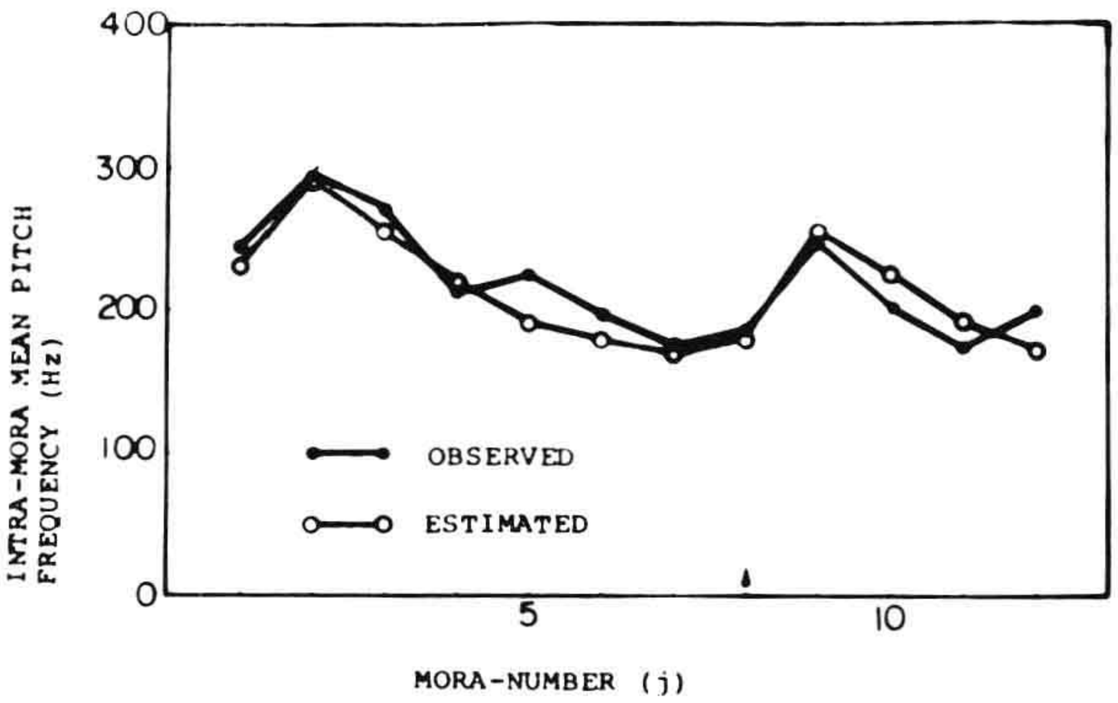


Fig. 3. Comparison of observed and estimated intra-mora mean pitch frequencies for the twelve-mora-sentence: "yūyake-zorani hoshiga-deta". The arrow on the abscissa indicates the first mora of the second segment.

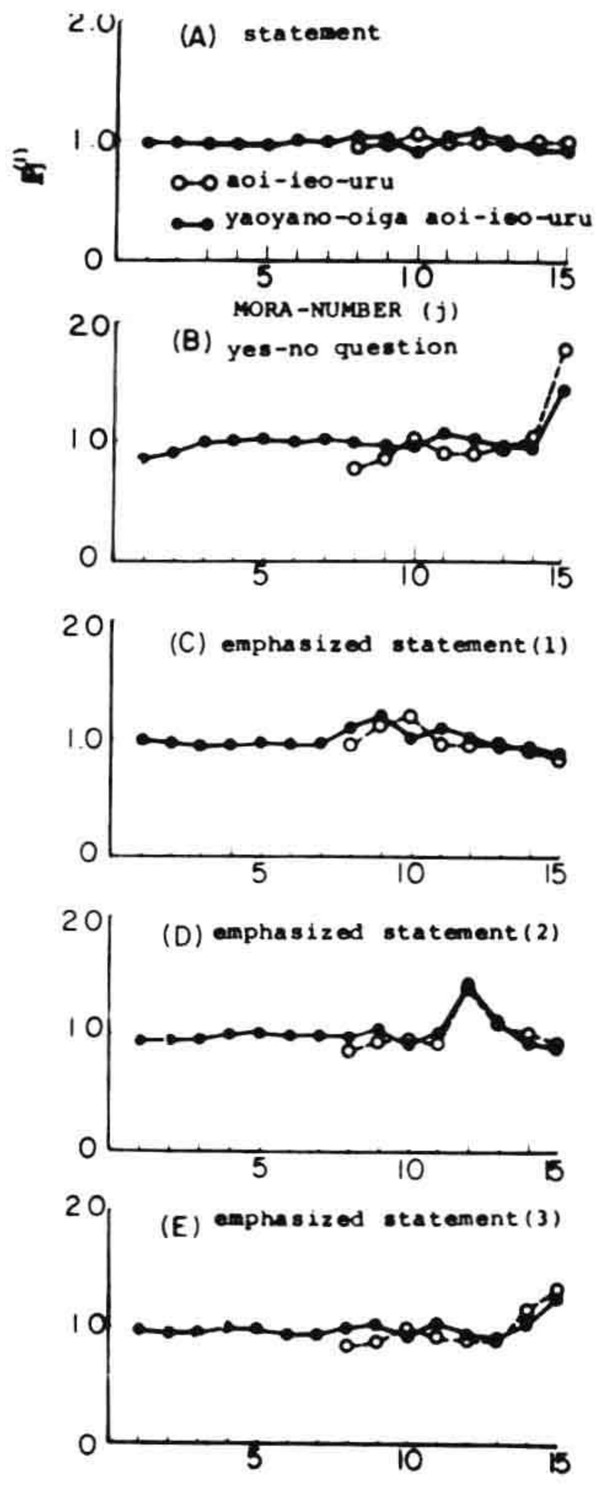


Fig. 4. Comparison of $P_j^{(1)}$ for longer and shorter sentences.

Reference

1) T. Nakayama et al., "Control Rule of Voice Source to Improve the Naturalness of Synthetic Speech", J.A.S.A., 42, 1162 (A) - 1967

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-6

Speech Synthesis by Rule

Akira Ichikawa and Kazuo Nakata

Central Research Laboratory, Hitachi, Ltd.

Introduction

To elucidate the coding process by which linguistic forms are transformed into speech sounds and to write the pronouncing thesaurus of audio response devices automatically, a rule is programmed to synthesize the continuous utterances of Japanese. When only orthographic symbols are used as input, there remain many unsolved linguistic problems. In this study, the input consists of semi-orthographic symbols with several special marks which designate the boundary of the accentual segment and that of breath-group, sentence style (e.g., declarative, interrogative), accent "fall" position, emphasis, and so on. The program has two modes; one is entirely by rule, and the other is partially by rule with extra data which provide the duration of each mora and buzz source data. It has a subroutine of vocal tract area function and a subroutine of calculation of formant frequencies, so it can control both vocal tract analogue synthesizer and terminal analogue synthesizer.

Duration of Phone

The strings of the input phonemic symbols are divided into mora, and then accent levels and durations of each phone are determined by our modification of Hiki's rule[1]. Examples of the durations of phones calculated by this rule are shown in Fig. 1.

Buzz Pitch Frequency and Intensity

The gross pitch pattern of one breath-group is calculated by the rule described in a separate paper[2]. The pattern is then modified by emphasis, phoneme, and accent level determined by accent "fall" position. The degree of these modifications decreases toward the end of each breath-group. The steep slope of pitch is regarded as the cue of the accent. The vowel following a voiceless stop has a descending pitch, starting at a higher frequency than normal.

The gross intensity pattern is approximated by two simple parabolic functions, rising and falling, because of its smaller contribution to the naturalness of synthetic Japanese[3]. It is, however, modified by phoneme, since the intensity is an important factor in certain phoneme identification.

An example of pitch pattern by the rule is shown in Fig. 2.

Vocal Tract Configuration and Formant Frequency

The vocal tract configurations of vovels, /h/, /y/, and /w/ are stored in the table named "phonemic table". Except for /h/, /y/, /w/, and /N/, those of consonants are calculated by a simple new co-articulation model. In this model the configuration of a consonant $S_{cv}(I)$ is derived as a multiplication of the individual co-articulation factor of the consonant $S_c(I)$ and the characteristic area function $S_v(I)$ determined by the sort of the succeeding vowel;

$$S_{cv}(I) = S_v(I) * S_c(I)$$

where I is the number of the section of the vocal tract. As for /a/, /o/, /u/, and /e/, $S_v(I)$ is an area function of /u/, and as for /i/, it is that of /i/. $S_c(I)$ is also stored in the "phonemic table".

On glide, the calculation of configuration has two modes: one is the simple linear interpolation at each section of the vocal tract and the other is the new model reported previously[4].

The target values of the formant frequencies are stored in the "phonemic table", and the linear interpolation method is adopted in transitional segment.

Some Japanese Phonological Phenomena Included in This Rule

The consonant preceding /i/ is palatalized commonly in Japanese. Contracted

consonant (expressed as /CY/ in Japanese, where /C/ means nonpalatalized consonant symbol) is treated by the rule of the co-articulation model mentioned above in the same way as the consonant preceding /i/. Duration, hiss source intensity, and off-gliding duration of these types of consonants are modified simultaneously by the rule. The assimilated consonant (symbolized first /C/ of /CCV/, where /C/ means consonant symbol and /V/, vowel symbol) is approximated by a pause of one mora length with succeeding consonant vocal tract configuration. The syllabic nasal /N/, a consonant peculiar to Japanese and forming one mora without vowel, is approximated by eight kinds of phones according to the succeeding phoneme with one mora duration. Devocalization, elision of vowel, descending diphthong, and relaxation are formed into the rule. One of the sonagrams of the synthesized speech is shown in Fig. 3.

Acknowledgement

The authors thank Dr. E. Matsui, Electrotechnical Laboratory, and Dr. S. Hiki,

Fig. -1

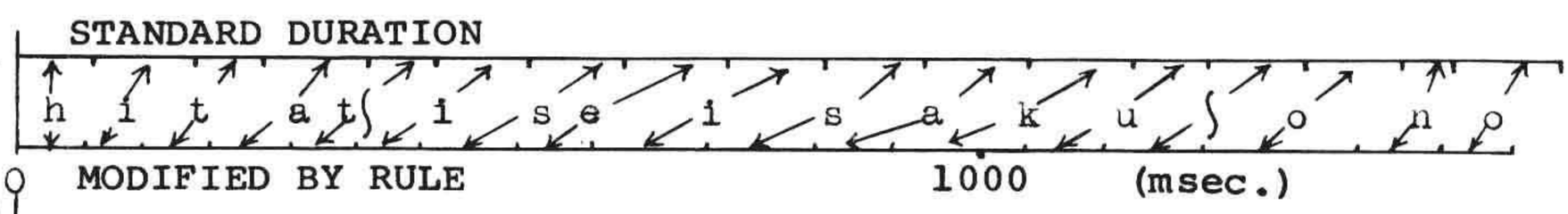
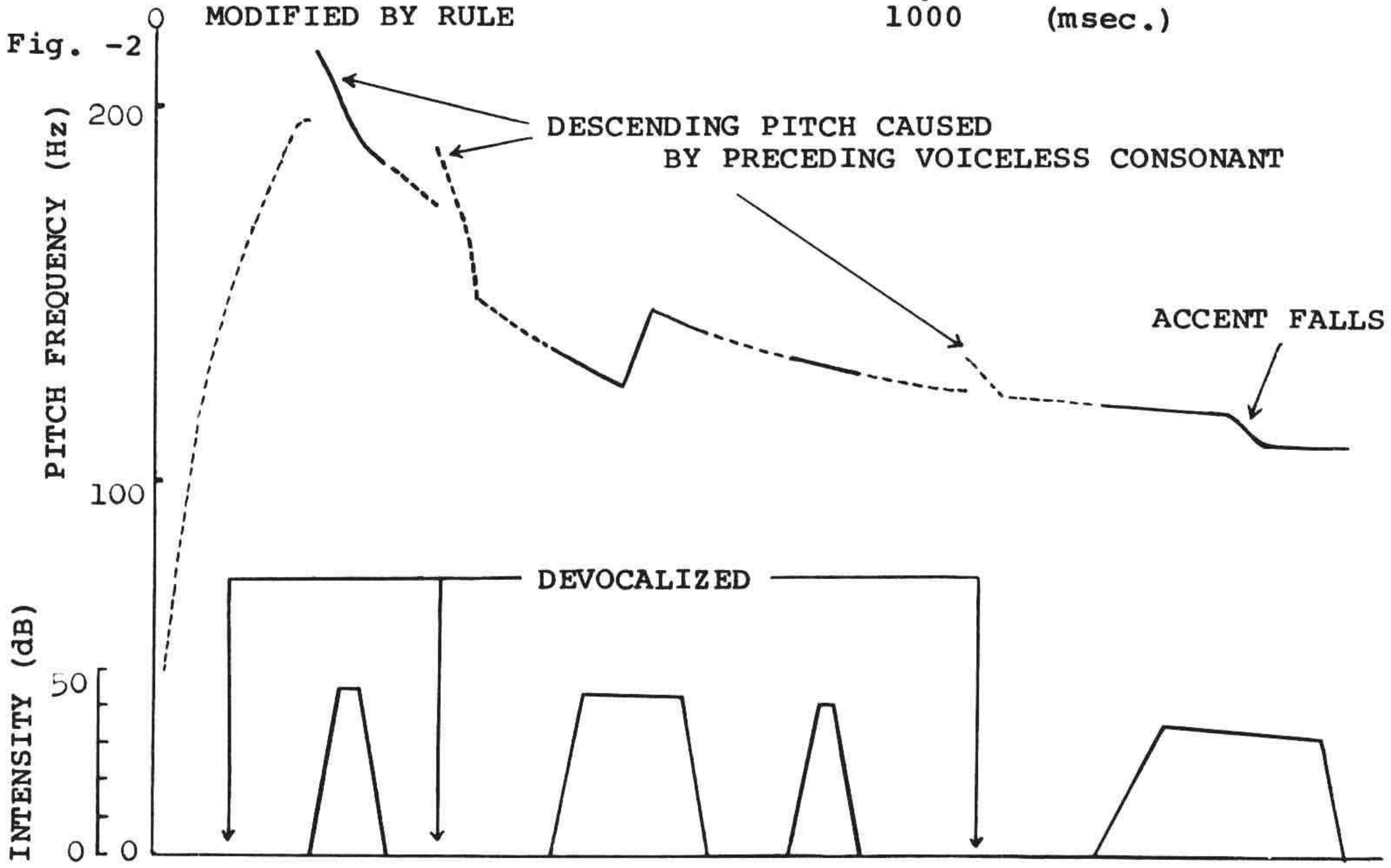


Fig. -2



Research Institute of Electrical Communication, Tohoku University, for their kind provision of the data on vocal tract area function and duration of phones.

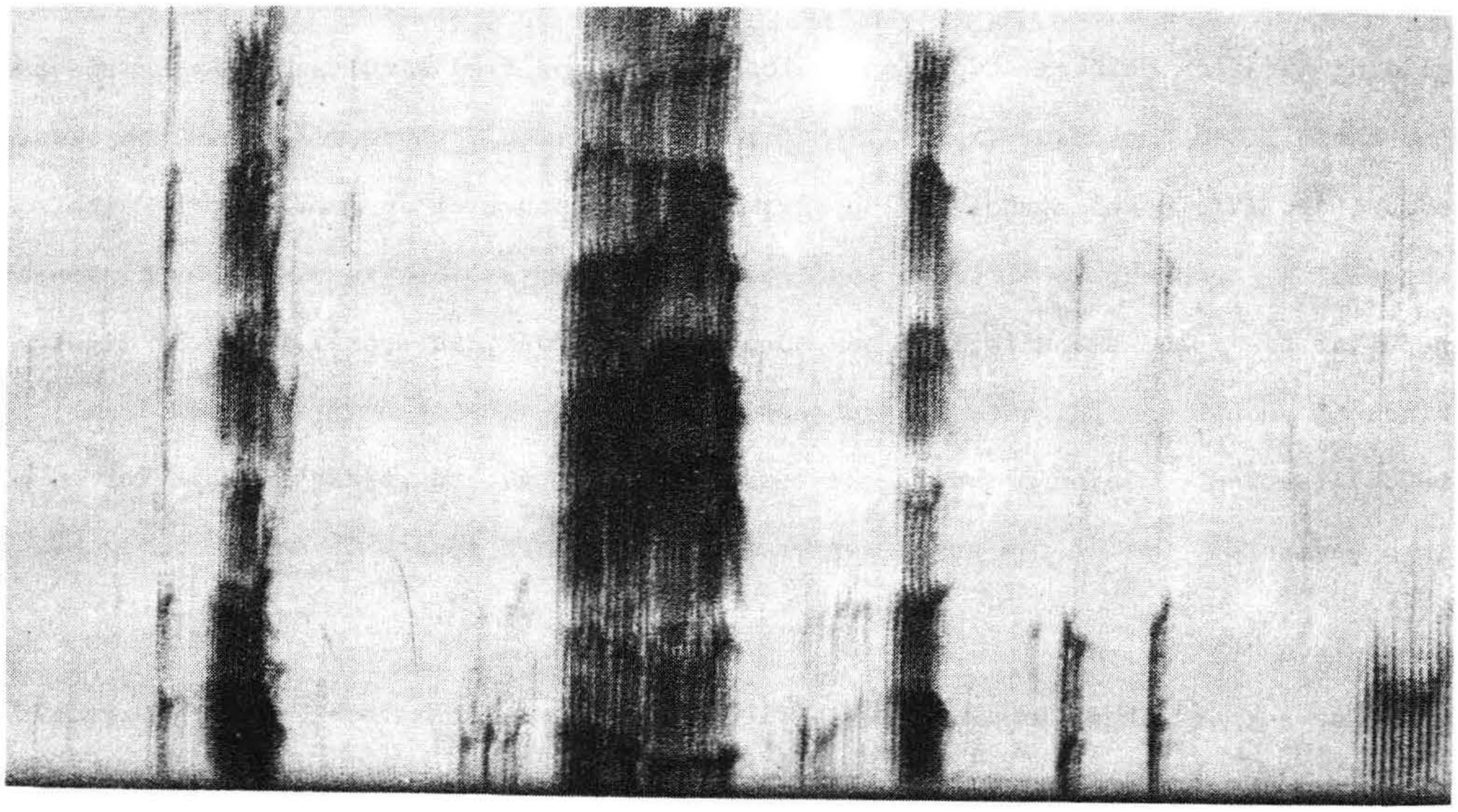


Fig. -3 A Spectrogram of Synthetic Speech
(Japanese /hitatʃiseisakuʃo/)

References

1) S. Hiki: Effect of the Context on the Duration of Phonemic Segment, J. Acoust. Soc. Japan, 23, 5, pp. 317-318 (1967).

2) T. Nakayama, A. Ichikawa and T. Miura: Generalization of Control Rule of Buzz Source to Improve the Naturalness of Synthetic Speech, The 6th International Congress on Acoustics, Tokyo, Japan, August 21-28, 1968.

3) T. Nakayama, A. Ichikawa and K. Nakata: Control Rule of Voice Source to Improve the Naturalness of Synthetic Speech, J.A.S.A., 42, 5, pp. 1163-1164 (A) (1967).

4) A. Ichikawa, Y. Nakano and K. Nakata: Control Rule of Vocal-Tract Configuration, J.A.S.A., 42, 5, p. 1163 (A) (1967).

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-7

A Hybrid Control System of a Human Vocal Tract Simulator

Takashi Honda, Seiichi Inoue, Yasuo Ogawa

The Faculty of Engineering, Meiji University

Introduction

This report describes about a speech synthesizer composed on a hybrid computer system that a digital computer controls a vocal tract simulator of a high speed analog computer. In order to synthesize speech in real time, and to control the simulator smoothly, high speed operational amplifiers[1] and high speed analog multipliers[2] are utilized for operational and controlling elements. The articulatory parameters needed to synthesize syllables are set up in control circuits and the informations sent to the control circuits from the digital computer are minimized, so that speech synthesis by means of on-line computer control may be realized easily.

Composition of the vocal tract simulator

It is well known that the vocal tract is regarded as a cascade connection of cylindrical segments of common length l_m but variable cross-sectional area A_m as shown in Fig. 1. Since each segment is equivalently represented by a series inductance L_m, a series resistance R_m and a shunt capacitance C_m, the cascade connection of such sections forms an equivalent circuit of the vocal tract. It can be shown that[3]:

$C_m = kA_m l_m/\rho c^2$, $L_m = \rho l_m/kA_m$, where ρ is the density of air, c the sound velocity, and k constant. The composition of an analog computer simulating the equivalent circuit is as shown in Fig. 2, where $\alpha_m = R_m/L_m$, $\beta_m = C_m/C_{m+1}$, $\gamma_m = 1/L_mC_m$. If α_m is constant, the configuration parameter to be controlled is only $\beta_m = A_m/A_{m+1}$ (m = 1, 2, ...), since γ_m is constant c^2/l_m^2.

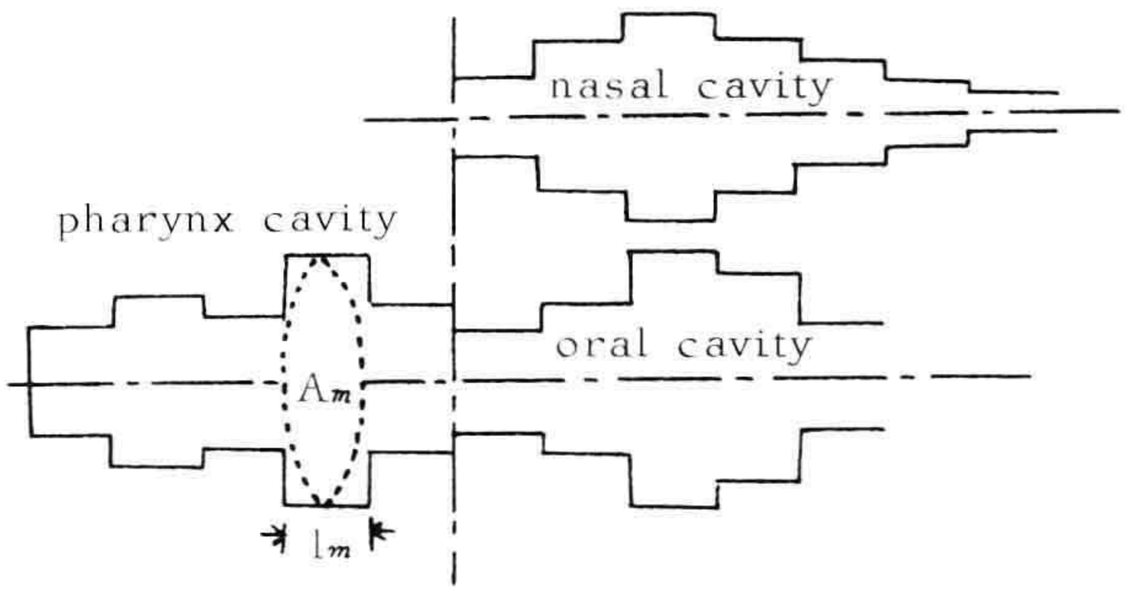


Fig. 1 Vocal tract approximated by the cascade connection of uniform acoustic tubes.

Here, the vocal tract is simulated by ten sectional segments, the maximum value of β_m is decided for 5 and the time constant of integrators is chosen for $1/\gamma_m^{\frac{1}{2}} = 5 \times 10^{-5}$ sec. under considerations of the available frequency band width of operational elements and of the error caused by approximation.

The nasal tract is simulated by seven segments and the junction point is laid between 5th and 6th sectional segment of the oral and pharynx cavity. The parameters controlling the nasal tract are $\beta_{n0} = A_{n1}/(A_{n1}+A_6)$ and $\beta_{n1} = A_{n1}/A_{n2}$, where A_{n1} and A_{n2} are the cross-sectional areas of 1st and 2nd segment of the nasal cavity respectively.

The radiation impeadance is approximated by a shunt circuit of R_0 and L_0, where $R_0 = 128\rho c/9\pi^2 A_{10}k$, $L_0 = 8\rho/3\pi^{\frac{3}{2}}A_{10}^{\frac{1}{2}}k$. The parameter controlling the radiation impeadance is only $\beta_{10} = \sqrt{A_{10}}$.

The principal merits of this composition are as follows: (1) This simulator can be controlled easily and smoothly by analog voltages, because analog multipliers are used for controlling elements. (2) Configuration parameters are very simple, that are A_m/A_{m+1}, $A_{n1}/(A_{n1}+A_6)$, A_{n1}/A_{n2}, $\sqrt{A_{10}}$. In this case, operational amplifiers of the frequency band width of 1MHz and analog multipliers of 250kHz were adopted.

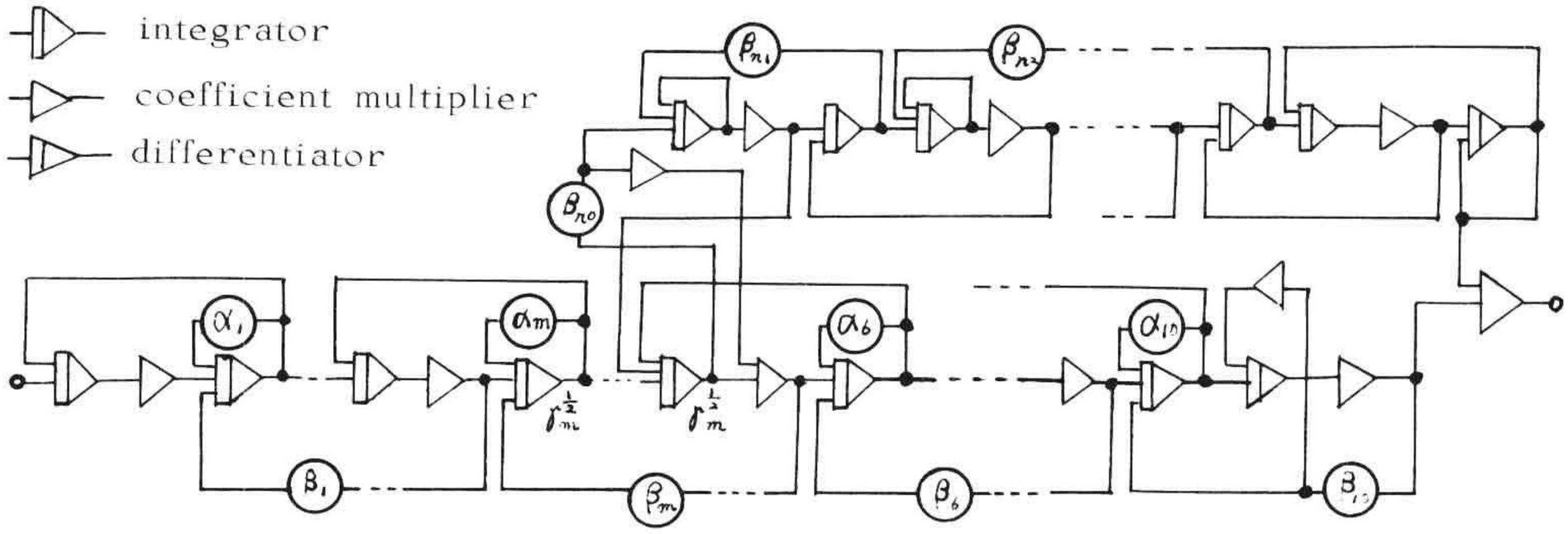


Fig. 2 Composition of the vocal tract
by means of operational amplifiers.

Control system and its operation

In order to make the control system simplify, following considerations are taken into: (1) As Japanese syllables are composed of about a hundred, continuous voices are produced by synthesizing each syllable one after another. (2) Many articulatory parameters used for synthesizing syllables are set up in the control circuits, and only five informations having relation to quality of continuous voices are given to the control circuits from the digital computer.

The blockdiagram of speech synthesis by means of this hybrid control system is as shown in Fig. 3. Operations are as follows: The first informations are sent to the configuration control circuit and the buzz and hiss generators, through the register, the decoder, and the D-A converters. The configuration control circuit has analog memories and generates analog voltages to control the simulator. The buzz and hiss generators also have analog memories and give the simulated acoustic sources to the simulator. Thus, one syllable is synthesized. The duration, the pitch, and the stress of the syllable are controlled by the output voltages of the D-A converters. The interval between one syllable and the next is kept by the syllable control circuit. Where, the information signals of one syllable are composed of twenty bits. Namely, seven bits of them are for the discrimination of a syllable, and thirteen bits of

them are for the control of the pitch, the duration, the stress, and the interval of the syllable.

Now, the articulatory parameters consist of configuration and acoustic source parameters. Seventeen kinds of configurations are provided to produce Japanese vowels and consonants, and ten potentiometers are use to set up the parameters of one configuration. The duration and the amplitude of buzz and hiss are controlled by modulators composed of monostable multivibrators and potentiometers, etc., in which acoustic source parameters are set up previously.

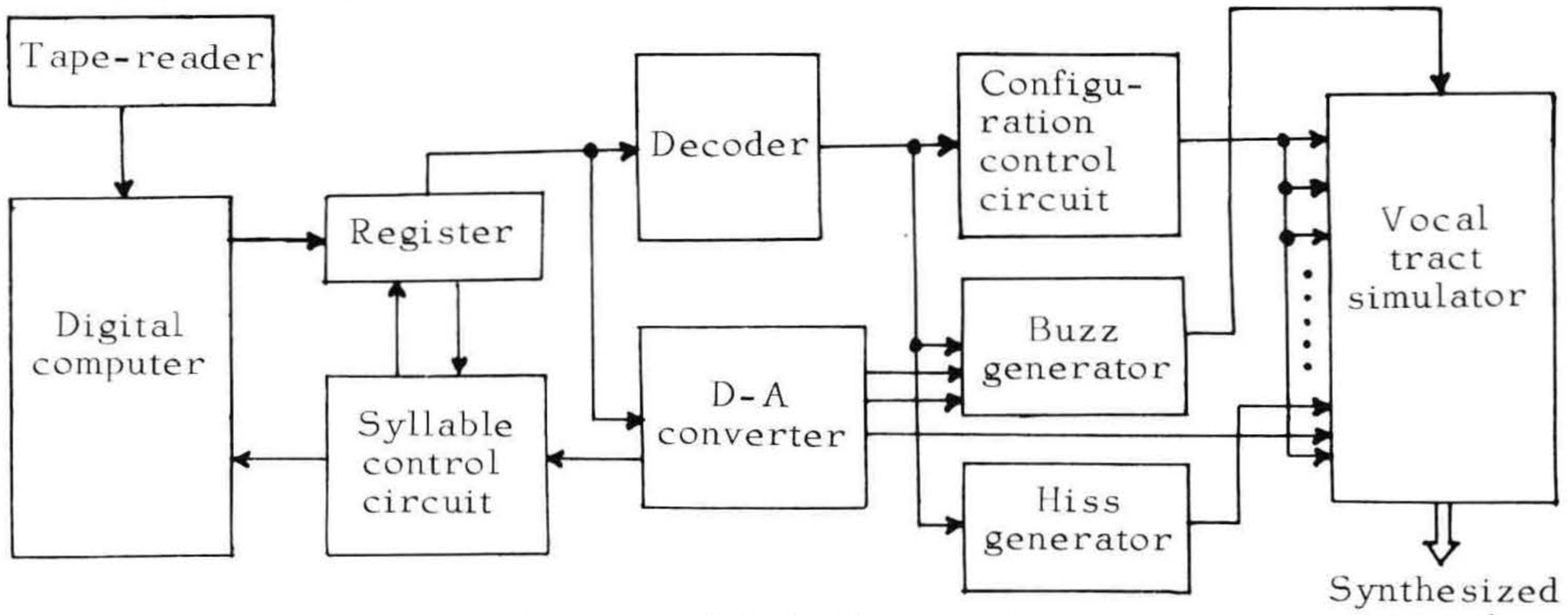


Fig. 3 Blockdiagram of hybrid control system

Conclusions

By the use of the high speed operational amplifiers to simulate a human vocal tract, the controlling parameters have been simplified, and with the high speed analog multipliers to control the parameters, the control system has been much improved. With these arrangements, the new type of speech synthesizer has been developed, which produces Japanese speech and songs audible enough continuously, with comparatively a few informations sent to the control device from the digital computer.

References

(1) Y. Ogawa, T. Honda: JIEE of Japan. Nov. 1965, p136.
(2) Y. Ogawa, T. Honda: JIEE of Japan, Jan. 1967, p216.
(3) G. Rosen: MIT Technical Report 353. Feb. 1960.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-8

Excitation of Vocal Tract Synthesizers

J. L. Flanagan and L. Landgraf
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey, USA

Simulation of the physiological processes of the speech mechanism promises a viable route to synthesis of speech. Toward this end, we have made a computer simulation of the vocal tract and the mechanisms of voiced and voiceless excitation. Control data required for synthesis are three physiological factors: subglottal pressure, vocal-cord tension and vocal tract shape.

Voiced Excitation

A network representation of the vocal system for voiced, nonnasal sounds is shown in Fig. 1. The tract is represented as a nonuniform transmission line whose sections represent plane-wave propagation in abutting right-circular cylindrical elements. The tract configuration is described by the cross-sectional area values (A_1 --- A_m). The radiation load at the mouth is Z_r and the acoustic driving-point impedance seen by the glottis is Z_t. The acoustic volume velocity through the mouth is U_m and that through the glottal port is U_g.

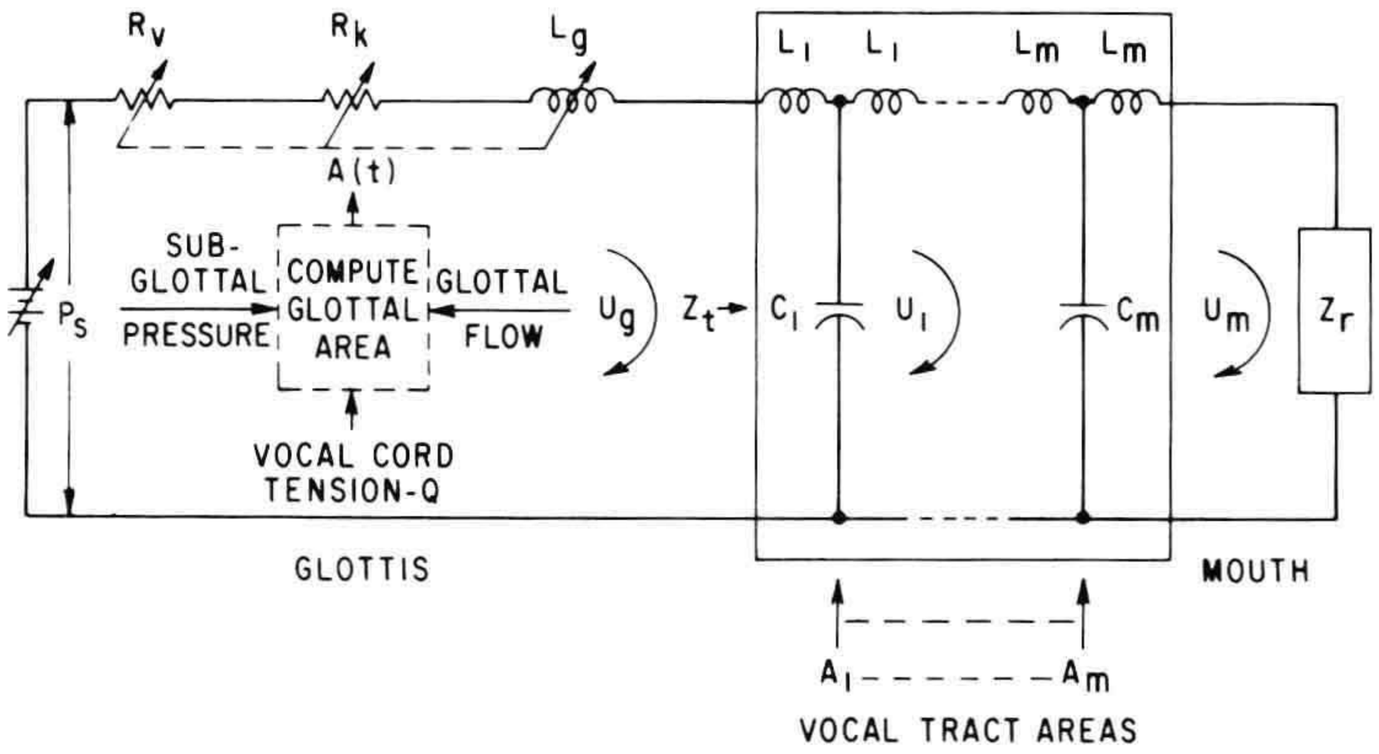


Fig. 1 - Network model for voiced excitation of the vocal tract

The acoustic impedance of the glottal orifice is represented by two loss elements, R_V and R_K, and an inertance, L_g. The values of these impedances depend upon the time varying glottal area $A_g(t)$. In addition, R_K is dependent upon $|U_g|$. The glottal area is linked to subglottal pressure, P_s, and to U_g through a differential equation that describes the vocal cord motion and its forcing function. Input data to the synthesizer are subglottal pressure, vocal-cord tension and the vocal-tract area function. From these physiological inputs, voiced sounds are synthesized which include irregularities in excitation waveform, pitch, and glottal duty factor, and which exhibit source-system interactions similar to those found in human speech.

The model treats the vibrating cords as a second-order mechanical system. The acousto-mechanical components are shown in Fig. 2. The cords have mass M, stiffness K and viscous loss B. At rest (x=0) their area of opening is the "phonation neutral" value A_{go}. The mechanical oscillator is forced by a linear combination of subglottal and Bernoulli pressures acting on the exposed intraglottal surface. The Bernoulli pressure is dependent upon $|Ug|^2$, which, in turn, is conditioned by the non-linear, time-varying glottal impedance. At complete closure, the glottal flow, and hence the Bernoulli pressure, becomes zero. Also upon closure, the contacting surfaces are assumed to be "sticky", and the oscillator damping is increased during the closure.

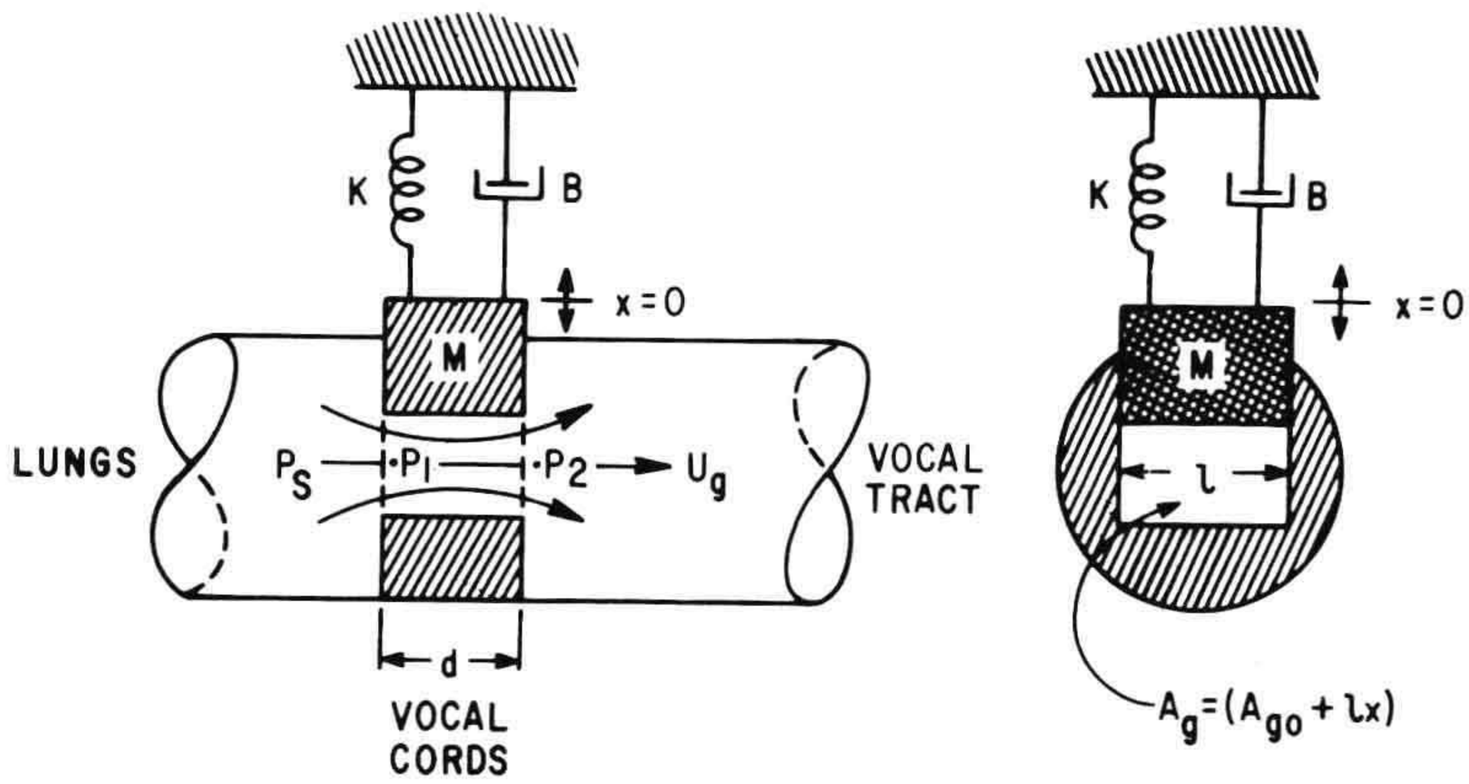


Fig. 2 - Representation of the vocal cords as a mechanical oscillator.

The continuous system shown in Fig. 1 can be represented by (m+1) differential equations which, in turn, can be approximated by difference equations. We have programmed these difference equations for simultaneous solution on a GE645 digital computer. The program accepts as input data time-varying samples of subglottal pressure P_s, cord tension Q and vocal tract areas (A_1 --- A_m), and computes sample values of all volume velocities, including the glottal flow and mouth output. The results can be plotted for visual display and D/A converted for audible output. A typical glottal area and volume velocity plotted by the computer for a vocal tract shape corresponding to /a/ is shown in Fig. 3. The figure shows the initial 50 msec of voicing.

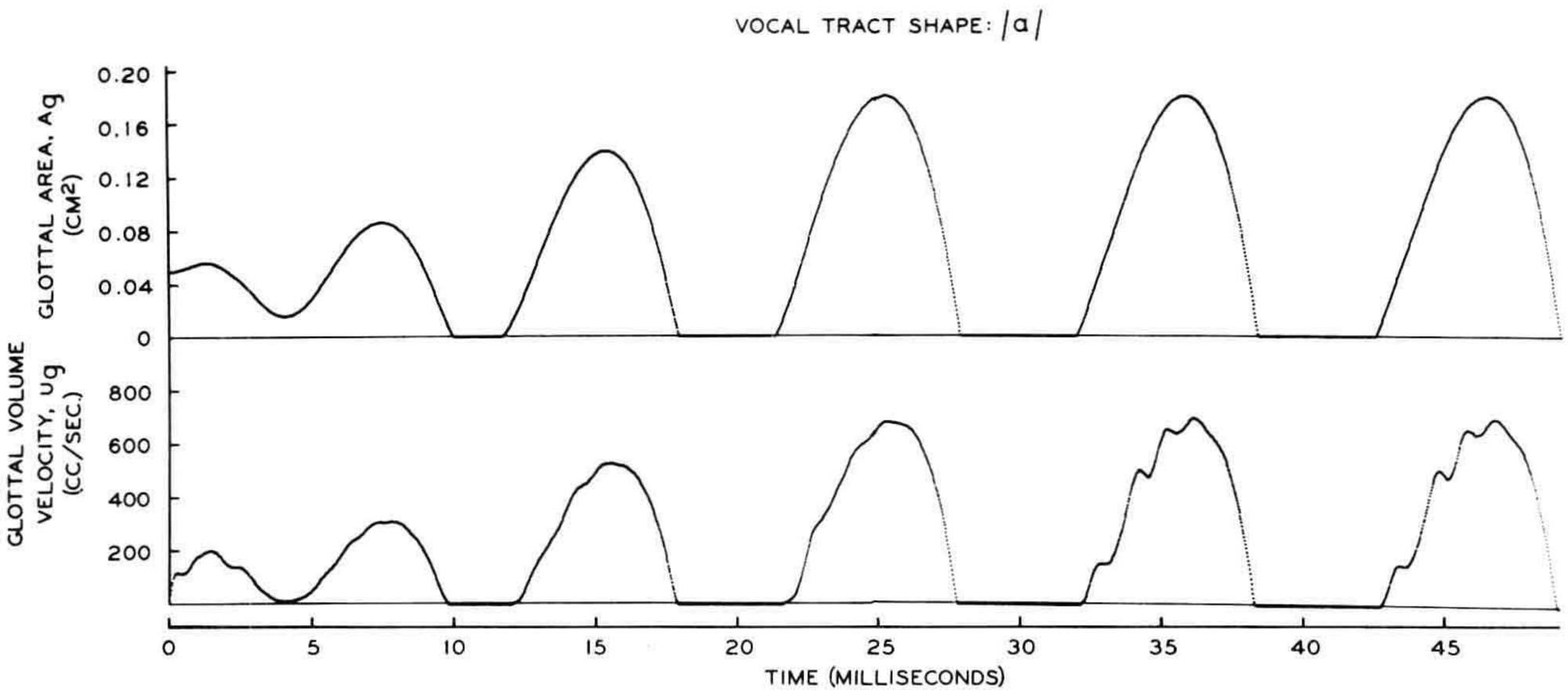


Fig. 3 - Computer plots of the calculated glottal area and acoustic volume velocity at the onset of voicing.

Voiceless Excitation

With slight modification, and with no additional control information, the physiological representation of Fig. 1 can be arranged to include voiceless excitation. Voiceless excitation can be of two kinds: continuant (fricative) and transient (stop). Fricative excitation (for sounds such as /s/ and /ʃ/) is generated by turbulent air flow at a constriction. The constriction is usually formed toward the front of the tract, and the vocal cords are relaxed and relatively open. The rib-cage muscles are contracted to raise the lung pressure and drive a flow through the constriction. At sufficiently high velocity (and hence Reynold's number), the flow becomes turbulent and constitutes a random, rotational sound pressure source.

Measurements indicate that, above a threshold value, the noise sound pressure generated by turbulence is proportional to the square of Reynold's number. The squared Reynold's number, in turn, depends directly upon the square of the flow velocity and inversely upon the constriction area. The effective noise source is spatially distributed, but generally it can be located at or immediately downstream (mouthward) of the closure. Measurements also suggest that its internal impedance is primarily resistive, and that it excites the vocal system as a serial pressure source. To first approximation, the noise spectrum is broadly peaked in mid-audio frequency range and falls off typically at about 6 dB/octave at low and high frequencies.

Stop excitation is produced by making a complete closure (usually labial, dental or palatal), building up pressure behind the occlusion and abruptly releasing it. A step function of pressure is therefore applied to the vocal network. The stop release is frequently followed by a fricative excitation (or aspiration) from turbulence generated after the release. This element of the stop is similar to a brief fricative continuant of the same articulation.

The transmission line vocal tract of Fig. 1 can be modified to approximate the non-linearities associated with turbulent flow. A serial noise source P_n, with internal resistance R_n, is introduced into each section of the vocal transmission line, as shown in Fig. 4. Both the level of the noise source and the value of its internal resistance are functions of the acoustic volume velocity and the area at that section. The noise level is modulated by a function proportional to the squared Reynold's number (in excess of a threshold value) and the source resistance is a flow-dependent loss, similar to the glottal resistance, and is determined by values of $|U_n|$ and A_n. By continually noting the magnitudes of branch volume currents and the corresponding cross-sectional areas, the synthesizer detects

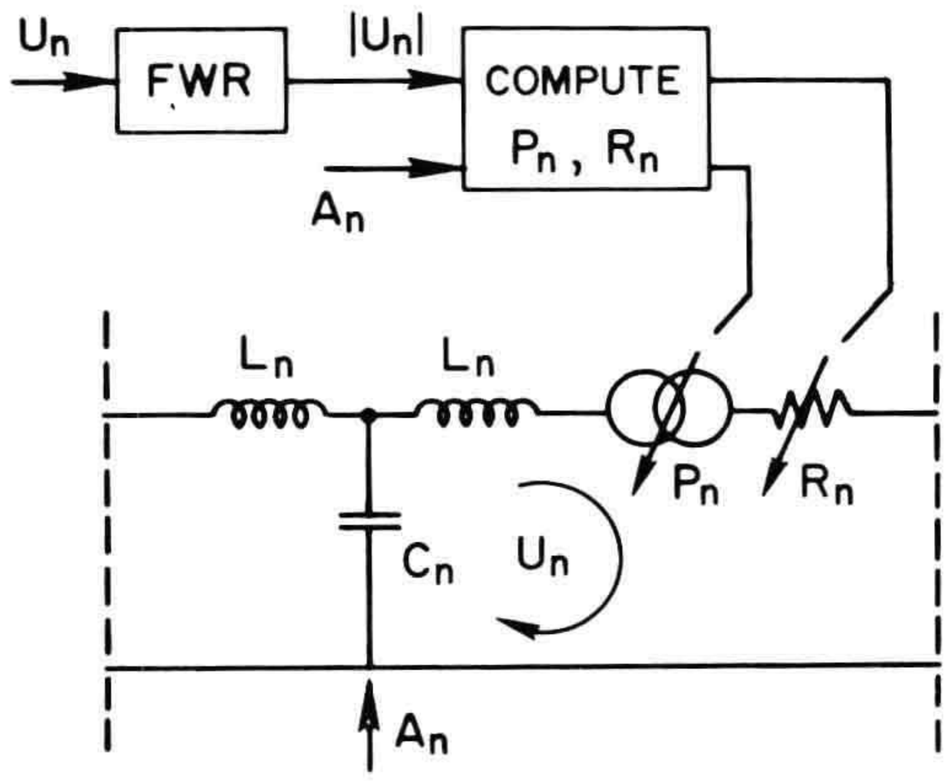


Fig. 4 - Automatic calculation and introduction of noise excitation at a constriction in the vocal tract.

conditions suitable to turbulent flow. Noise excitation and loss are consequently introduced automatically at any constriction. The control data (A_1 --- A_m), P_s and Q in effect determine the place of constriction and the loss and noise produced there.

The same operations generate voiceless stops, except in this case the control data produce a complete occlusion, a pressure build up, and an abrupt release. A pressure-step excitation of the tract and noise aspiration follow the release. Similarly, voiced fricatives are adequately simulated by simultaneous vocal cord vibration and noise generation. In this instance the tendency is for the noise to be generated in pitch-synchronous bursts, corresponding to the pulses of glottal volume flow during voicing. The result is a combined excitation in which the voicing and noise signals are multiplicatively related, as they are in the human.

References

1. J. L. Flanagan and L. Landgraf, Self-oscillating source for vocal-tract synthesizers, Trans. IEEE PGA and EA, March 1968.

2. J. L. Flanagan and D. I. S. Meinhart, Source-system interaction in the vocal tract, J. Acoust. Soc. Am., 36, 2001(A), 1964.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-9

A Model for Transitions of Vocal Tract Area Functions

Minoru Shigenaga, Hitoshi Ariizumi and Kiyoshi Tanaka

Faculty of Engineering, Yamanashi University

Kofu, Japan

1. Introduction

It is one of the fundamental problems in the study of speech to determine the vocal tract area functions and to investigate how they move during articulation. It is also important from the applicational points of view such as speech synthesis or recognition. Therefore first we have constructed a model of vocal tract configurations and their transitions for Japanese vowels. We have applied the model for vowels to consonants also by imposing proper distictive features. Though very simple, our models have given almost the same patterns of the first three formant frequencies as the observed ones.

2. Vocal tract area functions of Japanese vowels

We have determined the vocal tract area functions of Japanese vowels as shown in fig.1. Because the Japanese central vowel /u/ is apt to be influenced by adjacent sounds, we use two configurations $/u_1/$ and $/u_2/$. The lengths of the vocal tracts are all 17 cm, but we shorten these lengths by 8 %. However we reduce the degree of shortening as it approaches to /u/. After this we attach the lip whose cross-sectional area is equal to the area S(17) at x=17 cm (x is the distance from the glottis) and the length is M.

$$M = 0.46\{1+2.61\exp\{-0.024(S(17)-2)^2\}\cdot R\} \qquad (1)$$

where $R = 1/[1+1.5\exp\{-0.7(d-13.5)^2 - 0.83(Sd-0.625)^2\}]$

d = distance from the glottis to the place of articulation(cm)

Sd= cross-sectional area at d (cm^2).

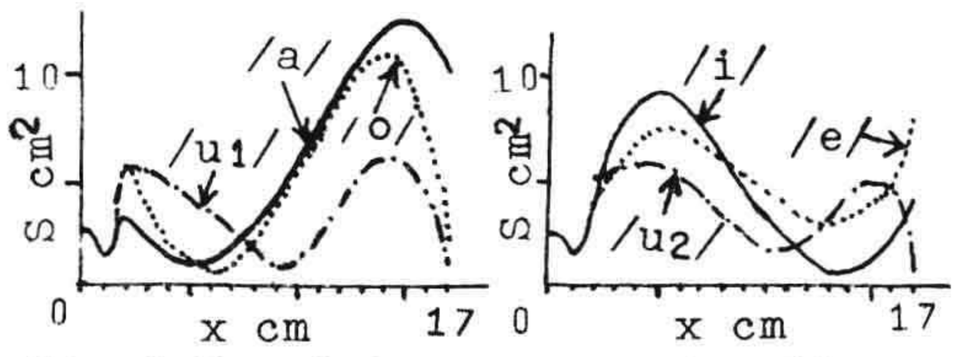


Fig.1 Vocal tract area functions.

Table. lp, lt = lip, total length

	d cm	Sd cm^2	F1 Hz	F2 Hz	F3 Hz	S(17) cm^2	lp cm	lt cm
i	13.5	0.625	290	2185	3050	4.1	0.9	16.5
e	13.0	3.25	490	1880	2680	8.1	1.0	16.6
a	5.5	1.25	730	1245	2650	10.0	0.7	16.4
o	6.0	0.625	480	845	2705	2.2	1.7	17.3
u1	9.0	0.8	315	835	2645	0.6	1.6	18.5
u2	10.0	1.8	305	1080	2340	0.6	1.6	18.5

For /i/ we have considered the influence of spread lip and compensated the length by R. The above table shows the formant frequencies and others calculated from the vocal tract configurations determined in such ways.

3. Transition of formant frequencies among vowels

Assuming that n=0 is the instant of leaving the vowel V_1 and n=1 the time of reaching to the vowel V_2, the average loci of F_2 (second formant frequency) drawn by the transition from V_1 to V_2 may be shown as follows. In the case of $F_2^1 > F_2^2$ (F_2^1 and F_2^2 are F_2 of V_1 and V_2.)

$$T_d(n) = \begin{cases} 1-\exp(-10.72n^3), & n < 0.4 \\ 1-\exp(-4.4n^2), & n \geq 0.4 \end{cases} \quad (2)$$

In the case of $F_2^1 < F_2^2$

$$T_u(n) = \begin{cases} 1-\exp(-4.4n^2), & n < 0.4 \\ 1-\exp(-10.72n^3), & n \geq 0.4 . \end{cases} \quad (3)$$

For successive transition among the three vowels V_1, V_2 and V_3 the locus of F_2 may be shown as the product of the function of the transition from V_1 to V_2 and the one from V_2 to V_3.

The above relations hold nearly good for the transition of the first formant frequency.

4. Transition of area function among vowels

$S_1(x)$ and $S_2(x)$ show the area functions of V_1 and V_2 respectively at x; S(x) the one during transition; T(n) the time function. If the following eq.(4) is assumed for transition of area function, the transition of F_2 depends almost linearly upon the variation of T(n)[1].

$$S^{1/p}(x) = S_1^{1/p}(x) + [S_2^{1/p}(x) - S_1^{1/p}(x)] \cdot T(n), \quad p=2\sim 3, \quad 0 \leq T(n) \leq 1. \quad (4)$$

In this paper we adopt 3 as p and use eq.(2) or (3) as T(n). For respective transition, it is necessary to modify the function and the following functions have been determined.

$$\left.\begin{aligned} &S^{1/3}(x)=S_1^{1/3}(x)+[\{K(n,x)\cdot S_2(x)\}^{1/3}-S_1^{1/3}(x)]\cdot T(n,x), \\ &T(n,x)=T(n)\cdot D(n,x)\cdot L(n,x)\cdot H(n,x), \quad K(n,x)=SL(n,x)\cdot E(n,x). \end{aligned}\right\} \quad (5)$$

Among these functions the more important ones are

$$\left.\begin{aligned} &D(n,x)=1-1.8Q_1\exp[14m^2/\{n(n-1)\}]\cdot\exp[-0.1(x-CX)^2] \\ &L(n,x)=1+1.25Q_2\exp[40m^2/\{n(n-1)\}]\cdot\exp[-0.2(x-16)^2] \\ &SL(n,x)=1-Q_3\exp[3(n-0.4)^2/\{n(n-1)\}]\cdot\exp[-0.5(x-CX)^2] \end{aligned}\right\} \quad (6)$$

$m=n-0.52$ for $F_2'<F_2^2$, $m=n-0.3$ for $F_2'>F_2^2$

CX=cross point between $S_1(x)$ and $S_2(x)$, lying nearly in the middle of the vocal tract. Q_1, Q_2 and Q_3 are constants.

When the three vowels V_1, V_2 and V_3 are uttered in succession, the transition of the area is expressed as follows:—

$$S^{1/3}(x)=S_1^{1/3}(x)+[S_2^{1/3}(x)+(S_3^{1/3}(x)-S_2^{1/3}(x))\cdot T(n-na)-S_1^{1/3}(x)]\cdot T(n), \quad (7)$$

$T(n)=0$ for $n\leq 0$, $T(n)=1$ for $n>1$

na=instant of beginning to carry out the command to move on to V_3 on the way of transition from V_1 to V_2.

5. Consonants

As well as vowels, we determine target configurations and move them by eqs.(5)~(7). But those equations must be modified a little, according to the characteristics of each consonant. The lip length is determined in consideration of the effect of rather spread lip.

Stop consonants[2]: The target configuration followed by a back vowel (suffix '1') and that followed by a front vowel(suffix '2') have been prepared for each stop consonant as shown in fig.2(same for both voiced and unvoiced). At the time of transition the negative areas are multiplied by the following function FF besides eqs.(5) and (6).

$$FF=\exp[-k|(n-n_0)/(1-n)|] \text{ for } n>n_0, \quad FF=1 \text{ for } n\leq n_0, \quad k=\text{constant} \quad (8)$$

The moment when the minimum area of S(x) reaches 0.1 cm^2 is considered as the instant of explosion and n_0 is the value of n at this moment. In the case of unvoiced stops, the effects of FF and L(n,x)

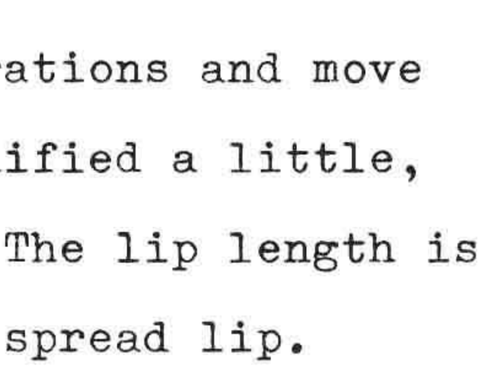


Fig.2 Some targets.

must be large. Fig.3 shows the transitions of formant frequencies of /aki/.

Fricative consonants : For fricatives it is sufficient to make the functions in eq.(6) operate over a rather wider range.

Nasal consonants[2]: It is sufficient for nasals to treat them in much the same way as we did voiced stops except for their coupling with the nasal cavity. The nasal cavity area function is varied within two sections(1 cm long) of the boundary between the nasal and oral cavities. It is taken into account that the degree of nasalization for the following vowel is more remarkable than that for the preceding one. Fig.4 shows the formant and anti-formant frequencies for /ima/.

For the other consonants the same treatments are available.

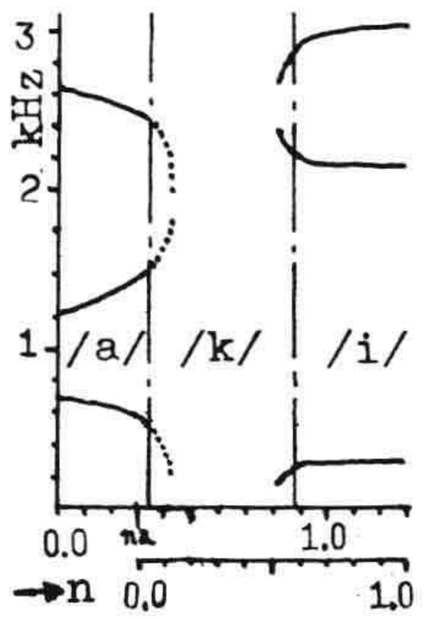


Fig.3 Calculated $F_1 \sim F_3$ for /aki/.

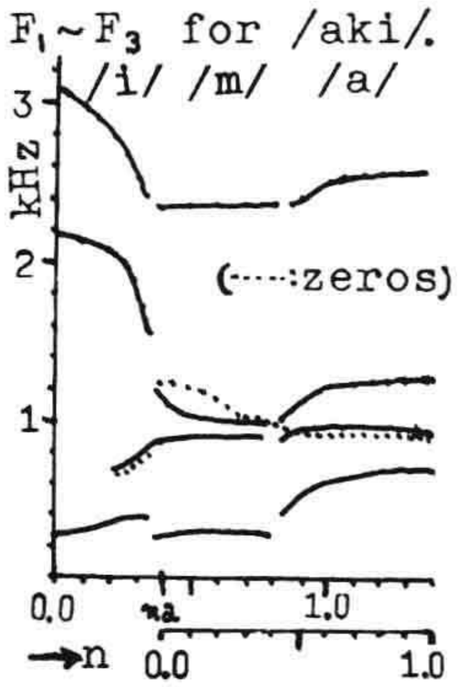


Fig.4 Calculated poles and zeros for /ima/.

6. Conclusion

We have determined the target configuration and the lengths of both lip and vocal tract of each Japanese sound uniquely.

For the transition of the area function between two vowels, if eq.(4) is used, the relation between the variation of T(n) and that of F_2 is almost linear, and for each case may be modified so as to realize the characteristic features. The transition among three or more sounds can be easily realized by eq.(7) or its expanded forms.

References

1. M.Shigenaga, H.Ariizumi and K.Tanaka, A model for transitions of vocal tract area functions for vowels. Reports of Committee of Automaton, Inst. Electronics and Comunication Engineers, Japan, Feb. 1968.
2. M.Shigenaga et al., Transitions of vocal tract area functions for stops (Nov.1967) and nasals (Apr.1968), Reports of Meeting, Acoust. Soc. Japan.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-10

A Computer Controlled Terminal Analog Speech Synthesizer

Takashi Hirasaki, Hikaru Date and Hirofumi Nakajima
NHK Technical Research Laboratories, Tokyo

Introduction

It is not necessary to point out the importance of a speech synthesizer to understand the nature of speech. A speech synthesizer of the terminal analog type has recently been constructed at the technical research laboratories of Japan Broadcasting Corp.. The synthesizer consists of a cascade-connection of formant circuits, anti-formant circuits, buzz-generator, hiss generator and some other circuits. All of these circuits are controlled by analog voltages which can be produced by either a digital computer or analog voltage generators.

In order to control the synthesizer by poles and zero in S-plane, a pair of conjugate poles(formant) or zeros(anti-formant) of the transfer function are realized in such a way that the two components of the pole or zero are controlled independently by respective analog voltages. The physical meaning of this is to control a resonant frequency and a half bandwidth independently by two different voltages.

In this paper, a new type of electronic attenuator, formant circuit, and anti-resonant circuit (consisting only of resisters and capacitors) are proposed. These were designed in accordance to the above considerations. An intensive investigation was performed also to obtain a wide dynamic range.

Formant Circuit

A transfer function of a formant circuit is realized by the circuit shown in Fig.1. The circuit in the broken line in the figure is a Miller amplifier whose gain is a function of the resistance, R2. The resonant frequency of the formant circuit is varied by changing the resistance. Half of the bandwidth of the formant circuit can be varied by resistor, R_σ . It should be noted that the resonant frequency and half of the bandwidth can be varied independently, provided that the square of the resonant angular frequency is much greater than the square of half of the bandwidth. In the actual formant circuit, both resistances R_2 and R_σ are replaced by the drain-source impedance of field effect transistors (FET) and the impedance can be varied by changing the voltage applied across the gate and source terminals.

Anti-formant Circuit

A transfer function of an anti-formant can be constructed by exchanging the numerator and the denominator of a formant transfer function. Since this function is unrealizable, two poles on the real axis are considered which are far away from the origin in comparison with those of the zeros. The new transfer function is given by

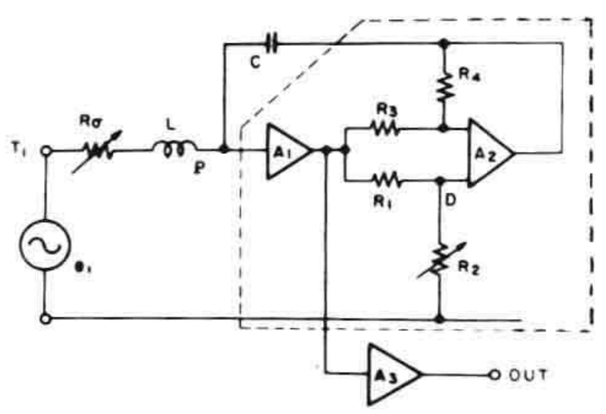


Fig. 1 Formant Circuit

$$T(S) = \frac{\delta_1\delta_2(S+\sigma_0+j\omega_0)(S+\sigma_0-j\omega_0)}{(\sigma_0^2+\omega_0^2)(S+\delta_1)(S+\delta_2)} = \frac{\delta_1\delta_2}{(\sigma_0^2+\omega_0^2)}\left[\frac{S^2}{(S+\delta_1)(S+\delta_2)} - \frac{(\sigma_0)2S}{(S+\delta_1)(S+\delta_2)} - \frac{\sigma_0^2+\omega_0^2}{(S+\delta_1)(S+\delta_2)}\right] \cdots (1)$$

where δ_1 and δ_2 represents the distance between the respective poles and the origin; σ_0 is the distance between the pole and the imaginary axis; ω_0 the distance between the pole and the real axis; and $S=\sigma+j\omega$.

The transfer function given by Eq.(1) excluding the coefficient can be realized by the circuit shown in Fig.2. It is easily understood that, assuming $\omega_0^2 \gg \sigma_0^2$, anti-resonant angular frequency ω_0 can be varied by changing the gain of the amplifier, G2, and half of the bandwidth, σ_0, by the gain of amplifier, G3. The gain of the amplifiers is varied easily by

exploiting FETs.

Electronic Attenuator

An electronic attenuator of a new type is proposed which consists of two transistors and an FET. The gain of a transistor with the emitter grounded is given by the ratio of the resistance connected to the collector terminal to that connected to the emitter, on condition that internal resistances of the transistor itself are neglected in comparison with the resistances connected outside. It is, therefore, easily understood that the gain of the transistor is varied by changing the value of emitter resistance.

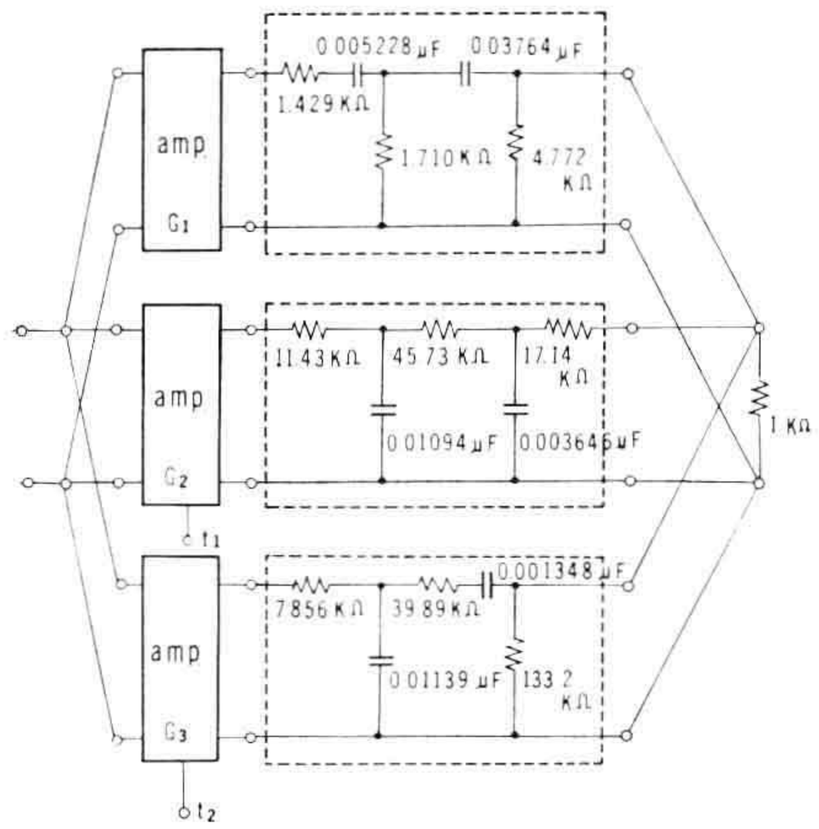


Fig. 2 Anti-Formant Circuit

It is a well known fact that the source-drain impedance of an FET is a function of the source-gate voltage and can be changed over a wide range(for example 10^3-$10^7\Omega$) by varing the voltage applied to the source-gate terminals under the condition that voltage applied to the source-drain terminals is less than about $0.1/2\sqrt{2}$ volts in r.m.s..

In the electronic attenuator proposed, an FET is employed as a variable resistance. This attenuator is shown in Fig.3. Transistor, Tr2 in the circuit supplies bias current to Tr1 whose gain can be varied.

An input signal is applied to the base of Tr1 and the output signal is taken out of the collector terminal of the same transistor. A capacitor inserted between Tr1 and the FET is for blocking the direct current of the FET flowing into transistor Tr1. Without the capacitor, the base line of the output signal will be modulated by the voltage applied to the gate-source terminals. The range of attenuation was about 60dB and this range depends entirely upon the FET used.

Dynamic Range of the Synthesizer

It was found that the noise level produced by the synthesizer agreed with the theoretical value. The only way left to increase the dynamic range would be to raise the level of the signal. According to the diagram shown in Fig.1, it is obvious that the signal passing through the formant circuit is applied to the drain-source terminals of the FET without attenuation, since the FET replaces R2 in the figure. It is very important to note that, at the resonant frequency, the input signal is amplified by factor Q and is then applied to the FET. This means that the maximum level of the input signal applied to the formant circuit is $0.1/(Q2\sqrt{2})$ r.m.s. volts. For example if the Q of the circuit is 30, this level will be 1.18mV. The dynamic range obtained in the circuit was about 50dB.

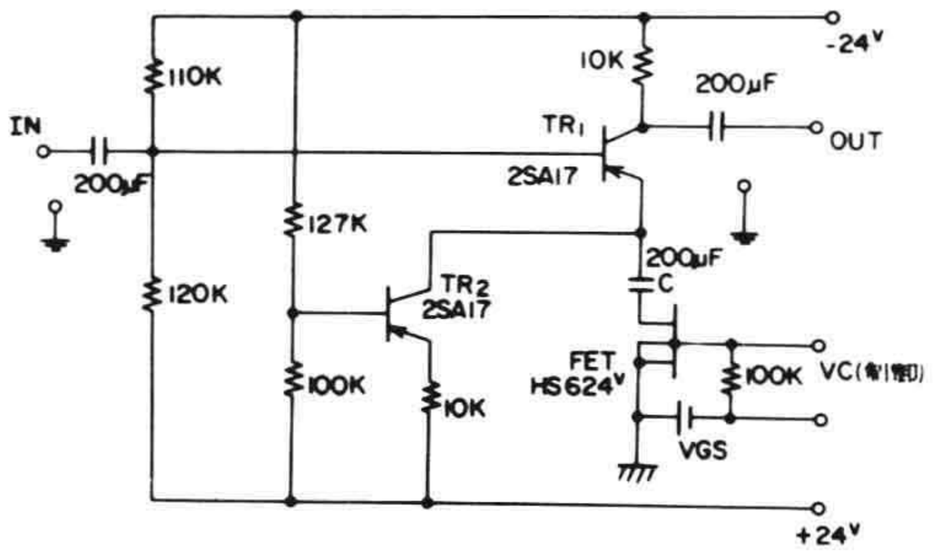


Fig. 3 Electronic Attenuator

The signal level was not be able to increase by inserting either a passive or an active transformer between the FET and point D in the circuit. The most effective way we found to increase the dynamic range was to apply feedback to the gate terminal of the FET from the source terminal. In this way about 20dB of the dynamic range for the formant circuit was increased resulting in 70dB in total. The dynamic range measured by a Bruel & Kjaer recorder is shown in Fig.4.

Conclusion

A solid state speech synthesizer with dynamic range of 70 dB was constructed.

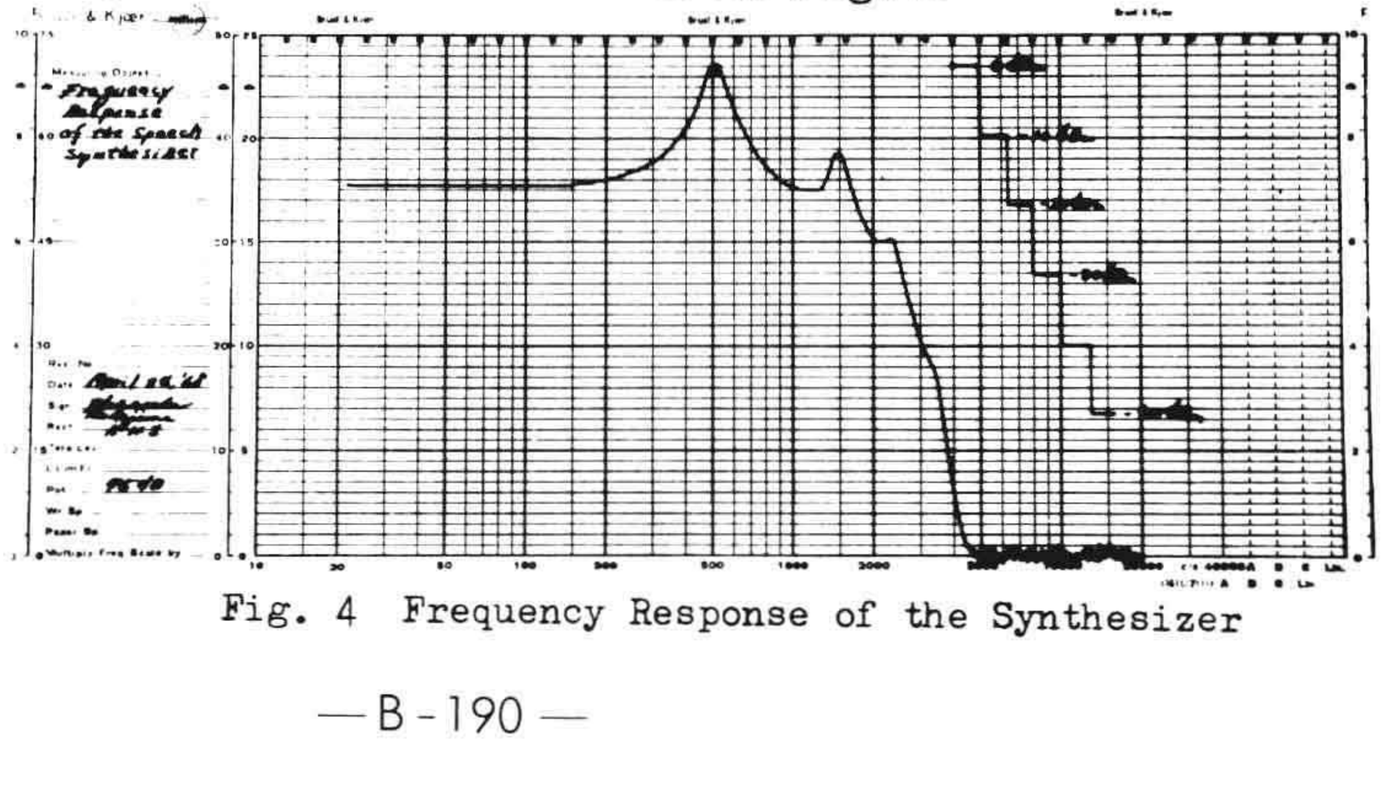


Fig. 4 Frequency Response of the Synthesizer

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-11

RECENT EXPERIMENTS IN PARAMETRIC SYNTHESIS OF POLISH SPEECH SOUNDS

JANUSZ KACPROWSKI, D.Sc. and WŁADYSŁAW MIKIEL, M.Sc.
Institute of Basic Technical Problems,
Polish Academy of Sciences,
Warsaw, Poland.

Introduction

The main objective of our actual research on speech synthesis is to provide general rules for the synthesis of current Polish speech in terms of a few parametric signals having a low information rate. The results of the present experiments should answer the following questions:

a. which acoustic parameters describing the speech signal are most important from the point of view of the perception of phonetic and linguistic information contained in it,

b. can these parameters be extracted from the natural speech signal when using reasonably simple technical means,

c. to which extent may the parametric signals, controlling the process of speech synthesis, be simplified, viz. quantized both in amplitude and in time, without noticeably affecting the intelligibility of the synthetic speech.

The answers to these questions not only throw the general illumination on the whole problem of speech synthesis itself, but also contribute to the development of new engineering systems for speech processing, coding and transmission, such as bandwidth compression and secrecy systems, automatic speech recognizers and many others.

Description of the Synthesizer

The resonance analogue speech synthesizer SYNFOR II used in the course of the present experiments has been developed from our former "terminal-analogue" synthesizer SYNFOR I reported previously [1] [2]. Although the general scheme of the instrument has been changed but slightly and resembles that one of OVE II [3], development has been made in several directions to improve its performance and flexibility for research purposes and to increase the precision of reproduction of data obtained from analysis. The most important modifications are the following.

The thyratron oscillator formerly used as the voice source has been replaced by a recently designed phantastron oscillator of high stability producing recurrent saw-tooth pulses. The envelope of the voice source spectrum decreases -6 db per octave and may be suitably shaped in two independent forming circuits placed in the vowel and nasal channel, respectively. The correction consists in successive integrating and differentiating the primary source function in simple RC networks. The resulting source spectrum envelope is thus flat up to the first pole frequency $s_1 = (R_1C_1)^{-1} = 2\pi \cdot 60\ \mathrm{sec}^{-1}$, then decreases -6 db/octave up to the second pole frequency $s_2 = (R_2C_2)^{-1}$ and -12 db/octave at higher frequencies.

The best results are achieved with $s_2 = 2\pi.2000$ sec^{-1} in the vowel channel and $s_2 = 2\pi.600$ sec^{-1} in the nasal channel.

The noise source produces a fairly uniform spectrum which, before being introduced into the consonant channel, is suitably shaped by low-pass and high-pass filtering. Both RC filters have -6 db/octave slopes and independently adjustable cut-off frequencies. In the synthesis of most consonants the noise is introduced also into the vowel channel after appropriate high-pass filtering which consists in decreasing the spectrum envelope -12 db/octave beneath 4000 cps for the sake of equalizing the formant levels in the low frequency range.

The attenuation in the four independently controlled amplitude modulators all using ECH 81 hexodes may be regulated within 64 db by varying their control voltages from 0 to -25 volts. To improve the accuracy of regulation in the low attenuation range, the modulators are so designed that the alteration of the control voltage from 0 to -12.5 volts decreases the gain from 0 to -15 db, whereas the remaining 45 db call for a further decrease of the control voltage from -12.5 to -25 volts. In each modulator an additional attenuation of 35 db may be introduced in 3 db steps to equalize the relative energy levels in all channels.

In our present synthesizer we do not use any zero-circuit either in the nasal or in the consonant channel. The lack of zero-circuit in the nasal channel results from our former research on the synthesis of Polish nasals [4]. The transfer function of the consonant channel, on the other hand, has only two poles, K1 and K2, the former with a high-pass and the latter with a low-pass characteristic. This simplification was acceptable since in the present synthesizer the consonant channel contributes only to the synthesis of the upper part of the spectrum, viz. above 2800 cps, whereas the lower part is synthesized in the vowel channel driven by the noise source and/or by the voice source. An apparent zero in the transfer function results, if necessary, from the summation of both parts of the spectrum. This fact is very convenient from the point of view of the design of a formant vocoder since the acoustic parameter denoting a zero cannot be easily extracted out of the natural speech signal.

The synthesizer as a whole may be driven by 11 voltage signals controlling the following parameters: the larynx tone frequency /F_0/ and amplitude in the vowel /A_V/ and nasal /A_N/ channel, the noise level in the vowel /A_H/ and in the consonant /A_C/ channel, and the pole frequencies in the vowel /F1, F2, F3/, nasal /N2/ and consonant /K1, K2/ channel.

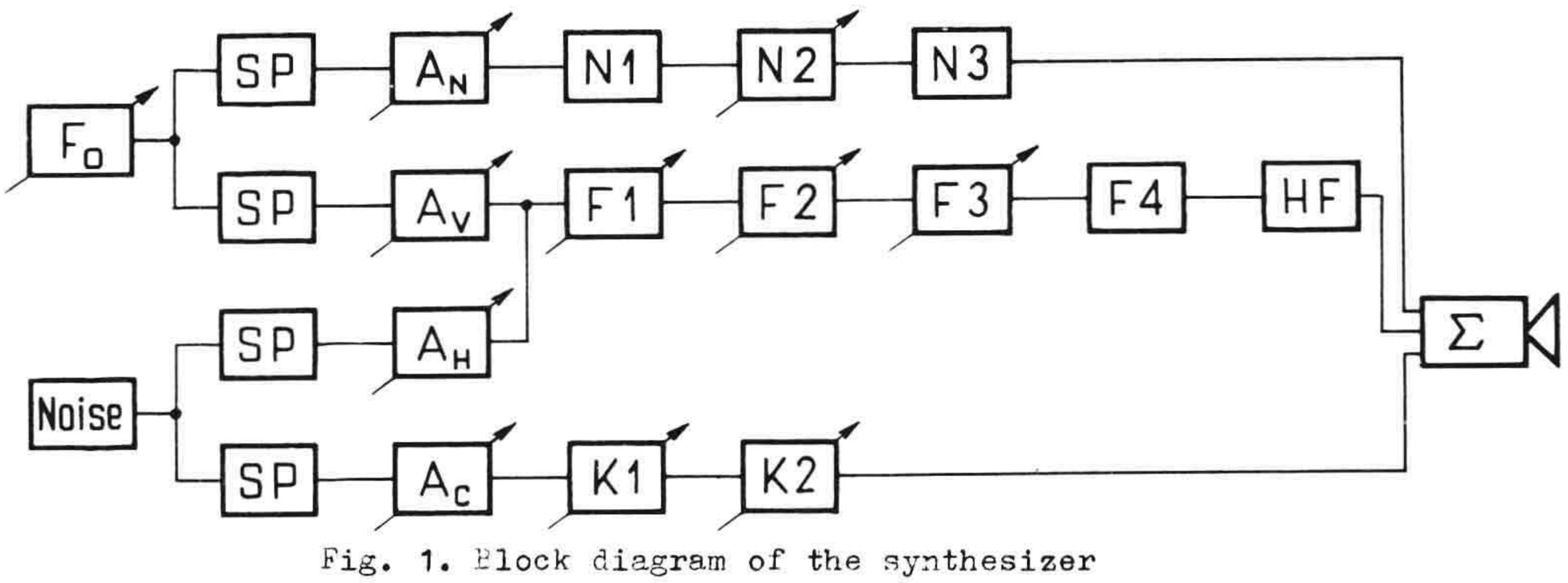


Fig. 1. Block diagram of the synthesizer

The control voltages, each ranging from 0 to 25 volts, are delivered by a function generator of the potential divider type. The control lines about 0.5 mm wide are drawn in conducting ink on a 30 mm wide plastic band which moves with constant speed v = 12.5 cm/sec and makes contact with a rotating cylindrical wire-wound potentiometer. The slowly varying control voltages are picked-up by a commutator divided into 12 segments. Every segment used for frequency control is 30 mm long, whereas every one for amplitude control is 16 mm long. Theoretically, the parametric signals used for pitch- and pole-frequency control may be quantized into about 32 levels, whereas those for amplitude control into eight 2 db

levels within the attenuation range from 0 to 16 db and into eight 6 db levels within the remaining range. Assuming that the parametric signals are sampled 40 times per second, we get the theoretical /maximum/ value of information rate equal to about 2000 bits/sec, which is in good agreement with practical demands for both formant vocoders and research purposes. The above data should be regarded, however, as merely indicative and must be lowered, as in fact some of the parameters, such as e.g. N2, K1 and K2, are quantized in no more than four levels only.

Experiments and Results

The part of our present program on speech synthesis to be reported here concerns the production of an inventory of nonsense VC and CV syllables, where V stands for each of the six Polish fundamentals vowels [a] [o] [u] [i] [ɨ] [e], respectively, and C denotes the semiconsonantal sounds [j] and [w] as in English words you and wool. In view of the constraints imposed by the structure of Polish language only the following connections occur in living speech:
VC syllables: [aj] [oj] [uj] [ij] [ɨj] [ej] and [aw] [ow] [uw] [iw] [ɨw] [ew],
CV syllables: [ja] [jo] [ju] [je] and [wa] [wo] [wu] [wɨ] [we].

All these syllables can be synthesized when using six parametric signals controlling the following parameters: A_V, F_0, F1, F2, F3 and event. F4. The timing patterns of the control signals were deduced by spectrographic analysis of the above given syllables uttered by male voices. The parameter A_V was measured by means of a logarythmic level recorder and evaluated as the over-all level of the speech signal corresponding to the syllable. The spectrographic data were approximated by continuous lines drawn on the moving band in the function generator. The criterion of success for the synthesis was not a faithful reproduction of the individual speakers' voice quality and dialect, but only the intelligibility and naturalness of the synthetic utterances. Informal listening tests showed that in spite of several radical simplifications in the timing patterns the quality of the synthetic syllables is quite satisfactory.

The synthetic syllables can be stylized as a sequence of three basic segments: steady state + transition + pattern at point of closest approach to target which corresponds to the final speech sound in each syllable. In CV contexts the initial steady-state segment assigned to the semiconsonantal part of the syllable was about 75 msec long and preceded a transition of 75 to 100 msec duration. Thus, at the start of every CV syllable the formant frequencies remained stationary at locus frequencies for about 75 msec, then moved on to the suitable vowel pattern and from there were stationary again for the next 150 msec terminating the utterance. In the VC contexts, on the other hand, the initial steady state was a little longer, equal to about 100 msec, whereas the character of the following transition depended on the terminating semiconsonant. In the case of the [Vw] contexts the transition lasted about 100 msec, terminated at formant frequencies corresponding to the vowel [u] and was followed by a steady state at these values. The only exception is the syllable [uw] which has no transitions at all. In [Vj] contexts the transitions were equal to about 150 msec tending to the points of closest approach to the formant frequencies of the vowel [i]. It should be mentioned that in this case the steady state does not exist, since the formant transitions are accompanied by the simultaneous decay of the larynx tone amplitude A_V, whereas the decay of voicing in [Vw] syllables occurs during the stationary segment corresponding to the vowel [u].

The listening tests showed that - in good agreement with the findings recently reported by Lindblom and Studdert-Kennedy [5] - the identity of the synthetic VC and CV syllables under investigation is determined not only by the formant frequency patterns at the points of closest approach to the target, that is to the initial and final steady state, but also by the direction and rate of adjacent formant transitions. It was proved that in this respect the rôle of F1 and F2 transitions is predominant. Formant F3 transitions contribute mainly to the naturalness of the synthetic syllable and may be varied within certain limits, viz. 10% for the [w] contexts and 20% for the [j] contexts without noticeably affecting the identity of the speech item. The same is truth for the fundamental fre-

quency F_0, except for the syllables [ij] [uw] and [wu] in which the formants are almost stationary. In this case the only cue for the recognition of the syllable is the alteration of the fundamental frequency.

To prove the above given general rules for the synthesis we have programmed and synthesized several phrases consisting of VC and CV syllables under investigation. Although such sentences, composed of vowels and semiconsonants only, are to some degree artifitial and appear rather uncommon in Polish, they were easily recognized and understood. An example of timing patterns and parameters values used for the synthesis of such a sentence is given in Fig.2. It follows that the general rules for the synthesis of isolated speech items, e.g. syllables, contain necessary information for programming longer segments of continuous speech, provided the mid-syllabic transitions are carefully taken into account.

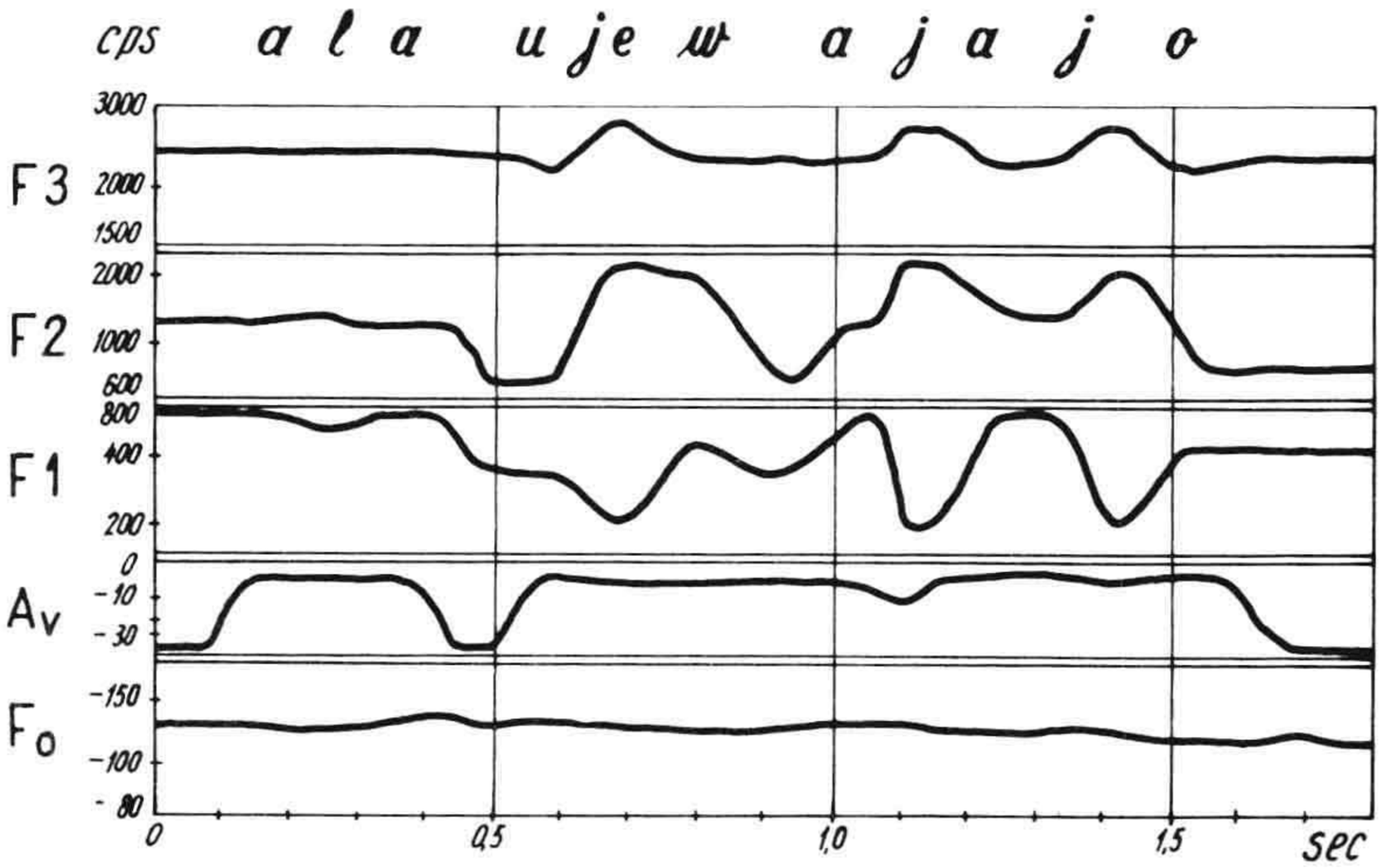


Fig. 2. The timing patterns of the synthesis parameters within the exemplary sentence

Acknowledgment

The authors are greatly indebted to Doc.Dr. Wiktor Jassem and his co-workers of the Acoustic Phonetics Laboratory for the spectrographic analysis of the investigated speech material and for their valuable help in adapting the spectrographic data for the synthesis process.

References

[1]. J.Kacprowski and W.Mikiel: Preliminary Synthesis of Polish Vowels by Means of Recurrently Impulsed Formant Filters. - Proc.Vibr.Probl., 4 (1963), 1 (14), pp.27-41.

[2]. J.Kacprowski and W.Mikiel: Simplified Rules for Parametric Synthesis of Nasal and Stop Consonants in C-V Syllables by Means of the Terminal-Analog Speech Synthesizer. - Acustica, 16 (1965/66), 6, pp.356-364.

[3]. G.Fant: Speech Analysis and Synthesis. - Royal Inst. of Techn., Speech Trans. Lab., Report No 26, June 1962.

[4]. J.Kacprowski: An Approach to the Synthesis of Polish Nasal Consonants by Means of Terminal-Analog Speech Synthesizer. - Proc.Vibr.Probl., 4 (1963), 3, pp. 235-254.

[5]. B.E.F.Lindblom and M.Studdert-Kennedy: On the Rôle of Formant Transitions in Vowel Recognition. - Journ.Acoust.Soc.Amer., 42 (1967), 4, pp.830-843.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

Speech Synthesis System Based on Interphoneme Transition Unit

Shuzo Saito and Shinichiro Hashimoto
Electrical Communication Laboratory, NTT
Tokyo, Japan

Introduction

This paper describes the speech synthesis computer system on the basis of vowel-consonant-vowel interphoneme transition unit[1](VCV concatination). The features of the system are that speech is synthesized on the basis of VCV concatination unit, that the input informations are composed of a series of discrete phonemic simbols and several tonal ones typewritten, that a terminal analogue synthesizer is simulated utilizing the on-line computer,[2] and that the devocalization rule[3]is supplied for obtaining more natural speech.

Hitherto phoneme[4]or syllable or diphon[5]unit has been applied to the speech synthesis system. However the rules for the synthsis have been complex, and that the synthesized speech by the rules has not always been satisfactory. The cause is mainly originated in that a phoneme behaves itself as if it were the other as the connecting object is replaced. In addition, not only the sort of a phoneme to be connected, but also the influence of toneme must be taken into consideration on the phoneme attitude. In case of connecting several phonemes together, it is therefore insufficient that acoustic features such as formants are only connected smoothly. If such relations pertaining to phoneme connection were all governed by rule, the rule would unnecessarily be highly complex. This may eventually come to describe every phoneme combination contained in a word orasentence.

Difficulty in the synthesis by rule lies mainly in the complicated connection at consonantal part. In order to simplify the rule, it is therefore preferable to connect phonemes in the middle of the vowel where it has comparatively gentle transition and longer duration. In Japanese words or sentences, except the initial one or two phonemes, it can be considered that almost all of them are composed of VCV or

VV phoneme sequences, for example "ohayoo" is composed of "oha", "ayo" and "oo". We call the acoustic features of the sequences "VCV or VV phoneme concatination" (in short "VCV concatination") and regard it as the synthetic unit. According to the idea, the difficulty described above will be much relieved. These VCV concatination data are stored into a certain memory in advance in the normalized form. In the speech synthesis procedure, they are arbitrarily read out of the memory and made into driving data and spectral data by comparatively simple rules.

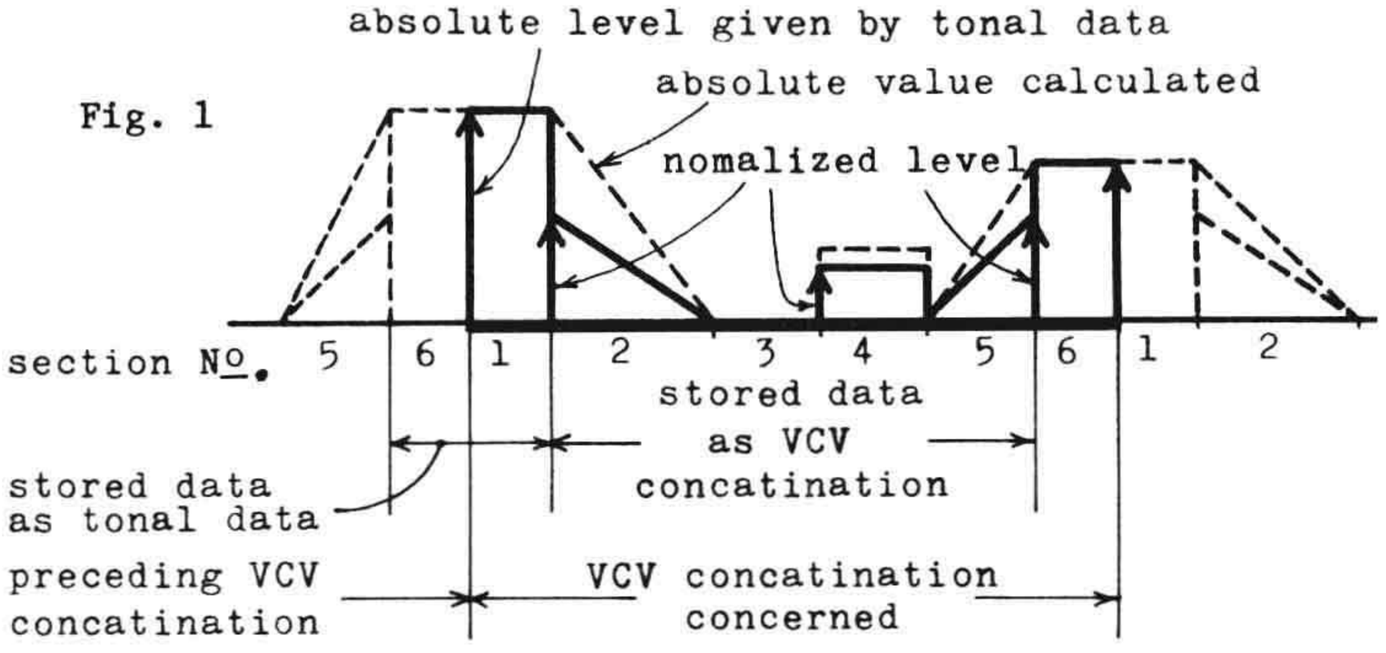
In taking these units into account, number of them comes to 900 including CV and V concatinations at the beginning of a word. There will be some possibilities to reduce this value from similarity of the units.

A Model of VCV Concatination

The model of the VCV concatination is composed of six sections as shown in Fig.1. The second, the third, the fourth and the fifth section of them correspond, as a whole, to a steep transitional part and each section has spectrum, level and time duration data normalized. They are stored as the VCV concatination data. The sixth section of the preceding concatination and the first section of the concerned one correspond to a gentle transitional part. The formant variation around here is made up by rule with the data of the fifth section of the preceding concatination and the second section of the concerned one. Although there are several rules for giving the formant variation, this model employs the simplest rule by which the final edge of the fifth section of the preceding concatination and the beginning edge of the second section of the concerned one is linearly connected. Level and duration characteristics in the sections are given by tonal data in absolute value. The absolute characteristics of the second and third section are therefore given in proportion to that of the first section, and the characteristics of the fourth and the fifth section are proportional to that of the sixth section.

It is already known that a word composed of n syllables has simply n accent patterns in Japanese.[6] Furthermore an accent pattern of a word can be approximated to those of the words with the accent kernels at the same syllable position, no matter how many syllables the words are composed of. Since Japanese words

Fig. 1



have at most 10 or 15 syllables, 15 sorts of tonal data corresponded to 15 accent patterns are prepared. The tonal data give fundamental frequency, level and duration time to each connecting point such as the sixth section of the preceding concatination and the first section of the concerned one. An example of tonal data of the word in which the accent kernel is at the first syllable is shown in Table 1. The relative fundamental pitch frequency inflection among the points where concatinations are connected is given by linear interpolation. Absolute fundamental frequency is calculated by the maximum fundamental frequency typewritten as a tonal simbol. The

Table 1

	Syllable No.					
	1	2	3	4	- - - -	15
Relative Fundamental Freq.	100	81	70	71	- - - -	65
Relative Level	100	88	37	26	- - - -	20
Relative Duration Time	28	48	45	45	- - - -	45

tempo of articulation could be adjusted by compression or extention of each section by simple rule. In order to improve naturalness of speech to be synthesized, the starting edge effect of the fundamental frequency could also be taken into account by rule to modify the value obtained by linear interpolation. An envelope of excitation source level can be selected on every section and the sort of phoneme. The rules for devocalization of vowel which play a great role in improving naturalness are formulated. These informations and parameters are contained in the VCV concatination data or in the synthesis program itself.

Synthesis Procedure and Its Result

From the standpoint as mentioned above, the speech synthesis program is carried out according as the flowchart shown in Fig. 2. A sequence of discrete phonemic and tonal simbol typewritten in the form, for example, "eNpi∘tsu.4,150" as a synthesizer input are at the first stage devided into VCV concatination units, where "N" denotes a syllabic nasal, "∘" does that the vowel preceding the simbol is devocalized, "4" means that the word has the accent kernel at the fourth syllable and "150" specifies the maximum fundamental frequency of the word. At the next stage, a sequence of VCV concatination data and tonal ones read out of the memory are, by rule, made up into driving source data and the spectral data setting a vocal tract characteristics. At the final stage, the model vocal tract is drived by the model source to obtain synthetic speech. The spectrograph of Japanese word "eNpi∘tsu" synthesized in this way is shown in Fig. 3a and the natural one is shown in Fig. 3b.

As the results, it is found that the speech synthesis on the basis of VCV phoneme concatination is useful to simplify the rules and to obtain better synthetic speech. As for synthesis of a sentence, the accent pattern specifying tonal information of a word should be superposed on a intonation pattern of the sentence.

Authers wish to aknowledge the invaluable advice of Dr. Oguchi, director of the Research Division of E. C. L., NTT.

References

1) S.Saito, S.Hashimoto and H.Wakita: Meeting of IECE of Japan, no.123, '66.

2) S.Saito and S.Hashimoto: Meeting of IECE of Japan,no. 2548, '68.

3) S.Saito and S.Hashimoto: Meeting of ASJ, no.1-2-14,p.69, April,'68.

4) A.M.Liberman,F.Ingemann, L.Lisker,P.Delattre and F.S. Kooper: JASA vol.31,p.1490,'59.

5) S.E.Estes,H.R.Kerby,H.D.! Maxey and R.M.Walker: IBM Journal, no.1,Jan.,'64.

6) S.Saito, S.Hashimoto and H.Wakita: Meeting of ASJ, no. 1.2-9, p.51, Nov.,'66.

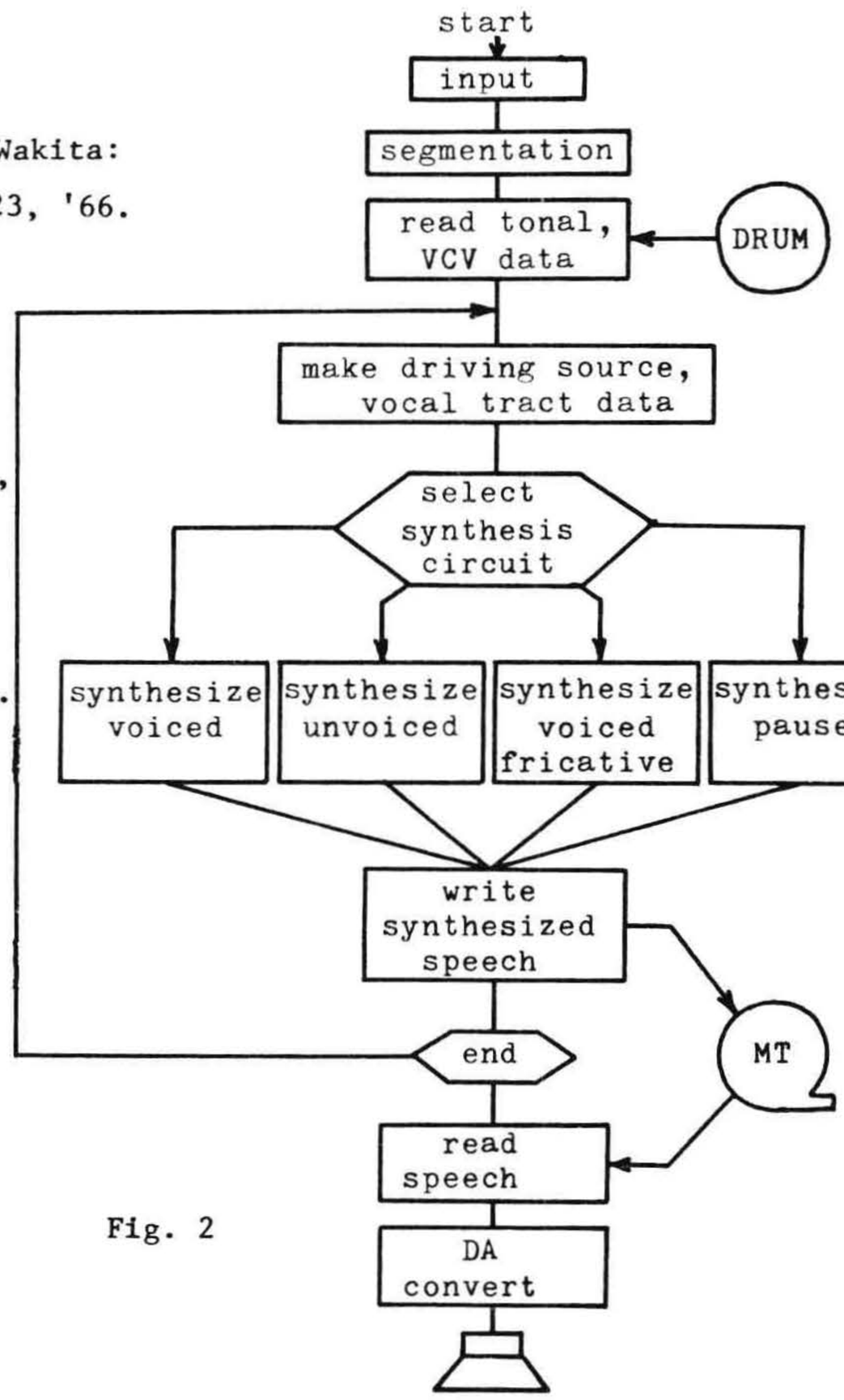


Fig. 2

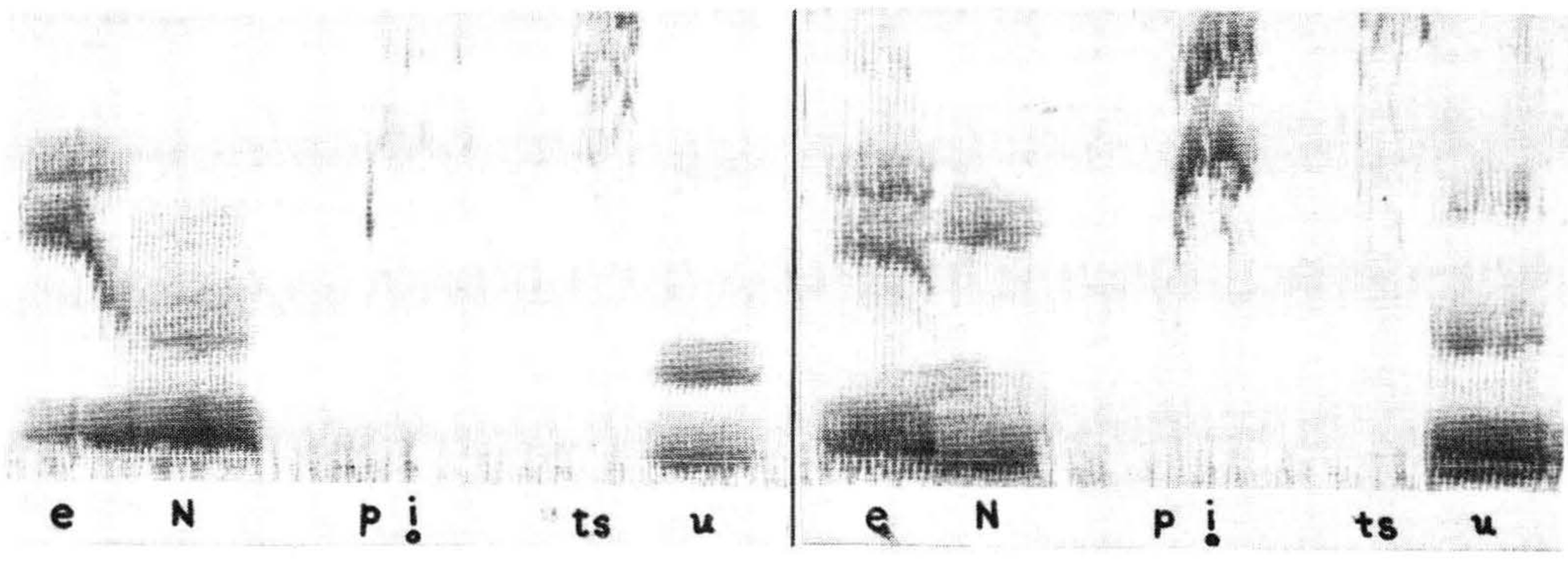


Fig. 3a

Fig. 3b

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-13

Speech Synthesis by Rule Supplementarily Using Natural Speech Segments

Yasuo Kato, Kazuo Ochiai, and Shintaro Azami

Central Research Laboratories
Nippon Electric Co.

From a practical point of view a terminal analog has attractive features, especially in its simplicity. Studies, however, on an optimization of a speech synthesis system with a terminal analog seem to be still in their early stage. In order to develope an efficient system of speech synthesis by rule for Japanese, we have conducted a series of perceptual experiments utilizing a flexible terminal analog synthesizer connected to a computer. Present paper describes several results of those studies and a synthesis rule which was consequently derived.

Experimental apparatus The terminal analog[1] consists of six variable pole circuits, two pole-zero pair circuits, a higher pole correction network, a radiation-shaping network, a spectrum tilting circuit, a noise generator, and nineteen controllable attenuators. Connections among these networks are easily changed on a patch board. A particular configuration of the hardware for our synthesis rule is shown in Fig. 1. In this scheme the unvoiced elements of the natural speech stored in a magnetic drum are used to reduce the system. Fig. 2 is a block diagram of the pole circuit. Real and imaginary frequencies of poles and zeros are digitally controled with the minimumn step of 1.5% change and with the maximumn rate of 10 k-times per second. High stability and wide range (F_1 : 70-2 kHz, $F_{2\sim4}$: 200-6kHz, $F_{5,6}$: 300-8kHz) of variation in pole frequencies are provided. Buzz source signals are generated in the digital computer NEAC-L3.

Preliminary perceptual studies 1)Formant transition function : As for the gradual transition from a vowel to an other vowel found in such Japanese words like /oioi/, /uiui/, etc., listeners could hardly perceive any degradation of synthesized speech quality when the transition functions were approximated by a single straight line[2]. Every synthesized word used here sounded so natural as listners could not separate the synthesized speech from the original human speech. Further, the results of informal listning test suggested that the situation for the rapid transitions to $F_{4\sim6}$ was useless for the naturalness of both vowels and consonants
2) Coarticulation : We were interested in to how much extent the coarticulatory deviations of formant loci contribute to the quality of speech synthesized by a terminal analog. We synthesized a number of words and sentences in which the vowels had the same formant frequencies regardless to the phonetic environments. They sounded quite natural. We have also conducted the following two sets of

experiments. First, VCV speeches (two V_S are identical) were synthesized and by listening tests, permissible and perceptible limits of F_2-F_3 loci of the consonant were obtained. An example of the results for /vgv/ (female voice) is shown in Fig. 3. Perceptible limits represented by solid lines are defined as the boundaries where a pair-comparison test (as-good-as test) gives 75% of preference to the standard items represented by dots in open circles. Items within the permissible limits, dotted lines, are firmly recognized as /vgv/ by all listners. There are no common permissible area to /aga/ and /igi/, and we can not neglect this kind of coarticulatory deviation of F_2-F_3 loci in the synthesis rule. Second, V_1CV_2 speeches ($V_1 \neq V_2$) were synthesized with various F_2-F_3 trajects. In Fig. 3, broken line A represents a F_2-F_3 traject of original human speech /agi/. Closing moment corresponds to the dot at the end of the arrow in the permissible region of /aga/, and opening time corresponds to the dot on the perceptible limit line of /igi/. In this example there is a large coarticulatory deviation of F_2-F_3 loci at the closing moment. Three kinds of synthesized speech /agi/, each of which has F_2-F_3 trajects represented A, B or C in Fig. 3, were presented to listners in pair, and comparison test was carried out. The results show that there are no perceptually meaningful differences among them. This kind of coarticulatory deviation seems to be unnecessary in a speech synthesis rule for a terminal analog. 3) Phoneme duration : In Japanese, a temporal unit mora, a term of prosody, is defined as a time duration which corresponds to that of a short monosyllable. Analysis data indicates that the mora duration deviates largely[3]. We supposed that the more speaker is trained so as to speak loyal Japanese, the less deviation will be observed. Therefore, the synthesized speech with constant mora will probably sound natural. We applied the constant mora rule to a number of words and to a few sentences. The synthesized speeches were quite natural. Durations of the consonants were predetermined. 4) Pitch frequency contour : By a series of perceptual experiments the following facts have come to light. a:-An effect of a rapid change of pitch frequency due to an abrupt change of an articulatory gesture, or due to the onset of voicing, can hardly be perceived. b :-Perceptual discriminabilities of the transitional speed and its timing of rise and fall in pitch frequency corresponding to a linguistic accent pattern are fairly low[4]. c:-A descent generally found in the pitch frequency contour over a breath group is perceptually meaningfull[4].

Formant pattern synthesis rule The input set of the rule is a linguistic description of the utterance containing phonemes and pause marks. As is shown in Fig. 4 a constant duration of 150 msec is given to every mora (C+V). The formant locus of a consonant F_{ic} or $F_{ic'}$ is given in a table. Formant transition time and duration of consonant are also given in the table. (Table 1.)

Pitch frequency contour generation Standard Japanese has pitch (tonal) accent, and in a word only a rapid fall (accent kernel) of the pitch frequency has a linguistic meaning. (In a connected speech a pitch rise plays a roll of denoting a word boundary.) We call here a word with an accent kernel type A, and the other without any accent kernel type B. Further, we classify type A into A_1 and A_2. Type A_1 has an accent kernel in the first mora, and type A_2 has an accent kernel in any other mora than the first. Summarizing the common features revealed in the analyzed data obtained from a set of carefully arranged sentences uttered by a lady radio announcer, and referring to the results of the preliminary experiments, the following rule was induced as a starting point. Breath group boundaries, word boundaries, pitch accent pattern and phoneme series (consequently mora series) are given to this rule as the input data. In Fig. 5 an example of a pitch frequency

contour generated by the rule is shown. The upper and the lower limits of pitch freqeuncy f_s and f_b are predetermined (for example f_s=320 Hz=2fb). The frequency f_i at the beginning of the breath group is determined by the relation $f_i = \sqrt{f_s \cdot f_b}$ (220 Hz in the example). A peak in the pitch frequency is put at the end of the first mora of each word if the word is type A_1, and at the end of the second mora if the word is type A_2 or type B. The peak value of the first word in a breath group is set at f_s and that of the last word is set at f_i. Between these two peaks a R-descent line is drawn as an envelope of the peaks of the other words. In a type A word, the pitch fall starts from the end of the mora where an accent kernel is placed. This fall continues over one or two morae until the pitch frequency gets to 65% of the peak value or to the lower limit f_b. After the fall terminates, the contour gradually goes down to the lower limit f_b until the following word appears. In a type B word, there is no such fall. In this case a small sink (Δf_{c1}, Δf_{c2} shown in Fig. 5) is given at the word boundary.

Unvoiced elements It is difficult to synthesize high quality unvoiced speech elements, maintaining the simplicity of a terminal analog. As described above, therefore, the stored unvoiced elements are used in our synthesis system. About twenty kinds of elements such as /s, ʃ , ss, ʃʃ , h, f, ts, k . . . . / are considered to be sufficient(5).

Quality of Synthesized Speech Speech thus far synthesized sounds quite natural. Over 80% of the articulation score for Japanese monosyllables has been obtained(6).

The authors are indebted to Dr. T. Sekimoto for his constant encouragement.

References

(1) Y. Kato and K. Ochiai ; Reports of the 1966 Autumn Meeting, The Acoust. Soc. Japan.
(2) K. Ochiai and Y. Kato ; Reports of the 1967 Spring Meeting, The Acoust. Soc. Japan.
(3) S. Hiki, Y. Kanamori and J. Oizum ; J. I. E. C. E., Japan, 50 (1967) 849.
(4) Y. Kato and K. Ochiai ; Reports of The 1967 Autumn Meeting, The Acoust. Soc. Japan.
(5) Y. Kato and K. Ochiai ; Reports of The 1967 Spring Meeting, The Acoust. Soc. Japan.
(6) K. Ochiai and Y. Kato ; Reports of The 1968 Spring Meeting, The Acoust. Soc. Japan.

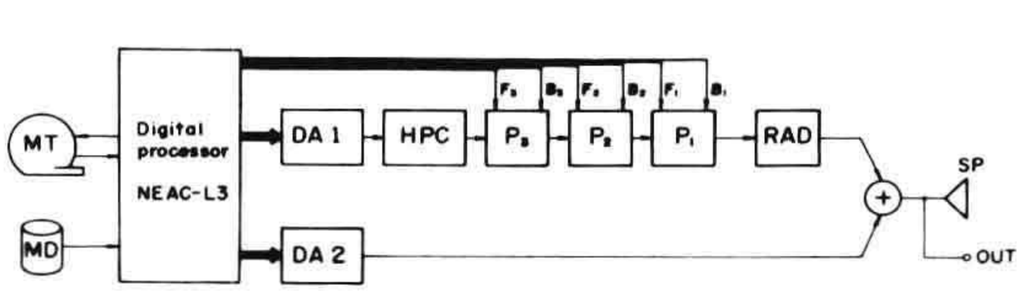


Fig. 1 Block diagram of a terminal analog speech synthesizer. The magnetic drum MD is used as a storage for the unvoiced natural speech segments such as [s], [ʃ] [f], [h], etc.

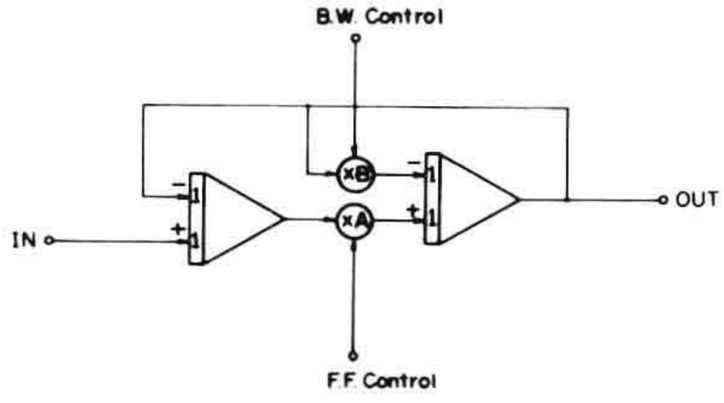


Fig. 2 Block diagram of a pole circuit.

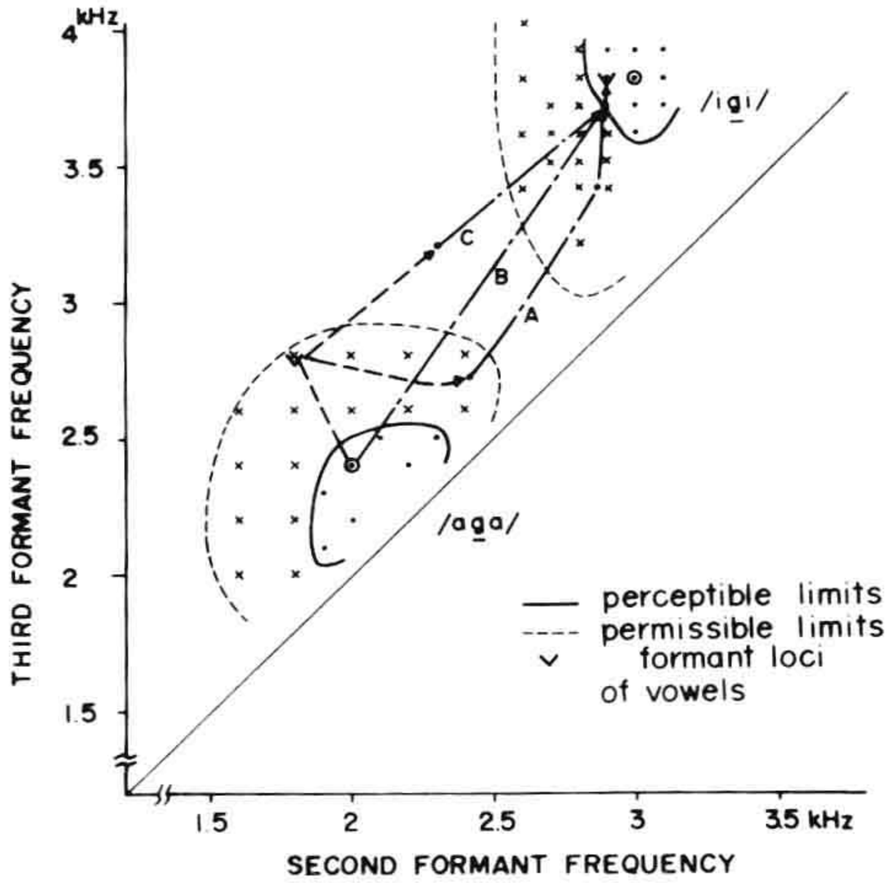


Fig. 3 Perceptible and permissible limits of F_2-F_3 loci of /g/. Lines A, B, C represent formant trajects of /agi/. (female).

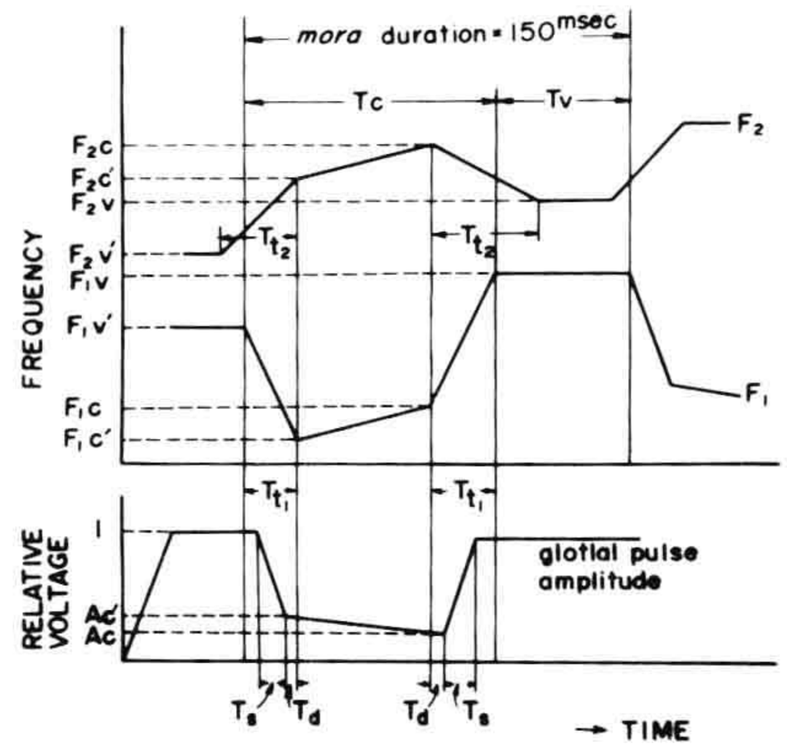


Fig. 4 Temporal patterns of formant frequency and of glottal pulse amplitude

Phoneme	Formant							Buzz Source			
	F_1 Hz	F_2 Hz	F_3 Hz	T_{C1} msec	$T_{C2,3}$ msec	T_{t1} msec	$T_{t2,3}$ msec	A_{C1}	A_C	T_d msec	T_s msec
b	200	1300	2100	80	130	25	50	0.3	0.3	0	20
d	200	1700	2800	90	90	30	30	0.3	0.1	0	25
g	200	2200	2400	130	150	40	50	0.2	0.2	0	30
r	400	1800	3200	75	75	25	25	0.5	0.5	0	25
w	400	1300	2900	100	140	40	60	1.0	1.0	0	30
m	300	1500	3200	80	90	15	20	0.8	0.8	0	15
n	300	2400	3400	70	80	15	20	0.8	0.8	0	15
t	300	2200	3000	110	130	30	40	0	0	10	10
k	200	1800	2900	110	130	40	50	0	0	20	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1 Part of data table which specifies temporal pattern of formant frequencies and of buzz-source amplitude. Fi varies in accordance with an articuratory environment, or an adjacent vowel. Fi in this table are for vowel /a/. (female)

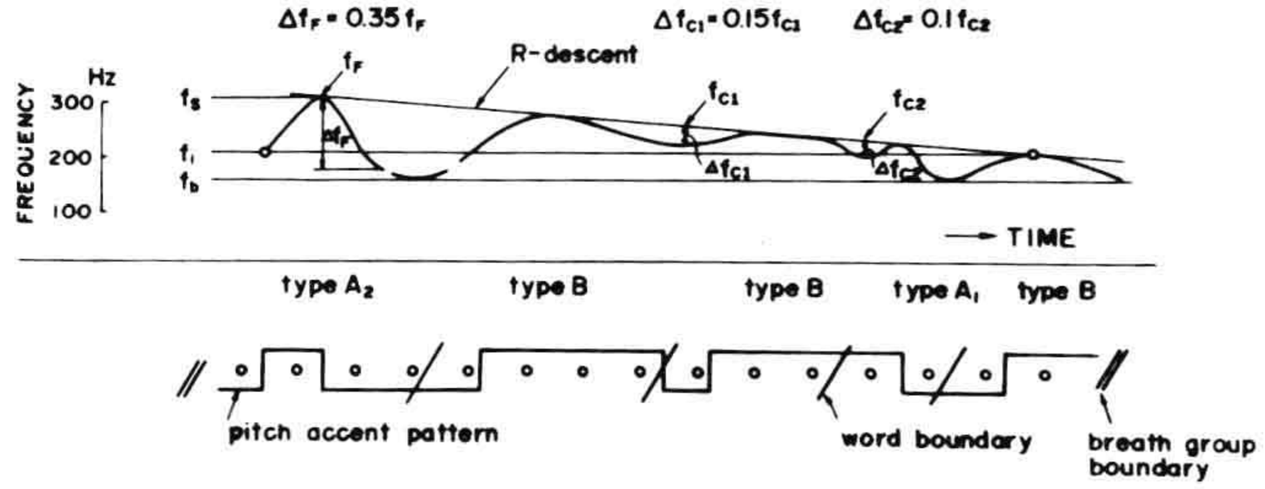


Fig. 5 Example of a pitch frequency contour generated by rule.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-14

New Results in Speech Synthesis by Rule

L. R. Rabiner and H. Levitt

Bell Telephone Laboratories, Incorporated

Introduction

In an earlier investigation,[1] a procedure for synthesizing speech from a discrete set of input symbols was described. In this method segmental data are calculated from a dynamic model of formant motion, and suprasegmental data from linguistic information (punctuation and stress marks) contained in the input string. A serial terminal analog synthesizer was used to generate the speech.

In the earlier investigation emphasis was placed on deriving rules for the segmental features and highly simplified rules were used for the suprasegmental features. This paper reports in part on an experiment designed to develop an improved set of rules for the suprasegmental features. Closely allied to this problem is the need for detailed data on fundamental frequency contours for real speech. Some preliminary results from such a study are also included.

Generation of Formant Data

Each phoneme in the input string is characterized by a set of formant target positions, and frequency regions surrounding the targets. The target positions contain quasi-static information about each phoneme - i.e. the steady state formant values of that sound. The frequency regions around the target contain information relevant to the dynamics of the formant transitions. These regions, along with the equations of formant motion, are used to account for such phenomena as vowel reduction,[2] coarticulation[3] and hysteresis of formant contours.

The formant contours are described by the solution to a second degree differential equation of the form:

$$\frac{d^2 f_N(t)}{dt^2} + \frac{2}{\tau_{AB}^N}\frac{df_N(t)}{dt} + \frac{1}{\left(\tau_{AB}^N\right)^2} f_N(t) = \frac{P_N(t)}{\left(\tau_{AB}^N\right)^2}$$

where: N is the formant number; (N=1, 2 or 3); $P_N(t)$ is a series of step functions describing the Nth formant target values for the string of phonemes; $f_N(t)$ is the response to $P_N(t)$ and corresponds to the contour for the Nth formant; τ_{AB}^N is the time constant of motion between phonemes A and B for the Nth formant.

The frequency regions represent the degree to which the set of formant values must approximate the target values of a given phoneme before initiating motion to the next phoneme.

A simplified example of formant motion (for a two formant system) is shown in figure 1. The frequency regions are indicated by Δ's and the initiation times of phoneme changes by t1, t2, and t3. Note that the time constants of the formant transitions are specified independently of each other.

Suprasegmental Rules

A physiological model[4,5] for generating fundamental frequency (Fo) contours has been assumed, consisting of two basic components - laryngeal tension (LT) and subglottal pressure (Ps). Each component is determined for a breath-group from stress and phoneme information contained in the input string. For most cases LT is fixed and Fo is linearly related to Ps.

The Ps contour consists of three principle components; an archetypal contour for an entire breath-group, a local perturbation due to unvoiced or voiced consonants, and a perturbation due to vowel stress. Figure 2 shows a typical Ps contour (or Fo contour assuming LT constant) and its composition from its component parts. The effect of vowel stress is manifested in at least two ways - fundamental frequency is increased and the vowel is lengthened in duration. In the earlier model a fixed increment in Fo and a near constant increment in duration was used for each stressed vowel. This procedure was considered inadequate and a new set of rules was developed allowing for four possible stress levels. The parameters of the model were the magnitude of the increase in Fo and the amount of lengthening of vowel duration, both as functions of the specified stress level. An increase in Fo also brought about a corresponding increase in speech intensity since pitch pulses of fixed energy were used.

A perceptual experiment was carried out in order to determine appropriate parameter values for the new model. Twelve versions of three separate sentences were generated for all possible combinations of: a) three nominal durations (50, 100 and 200 msec), b) two Fo increments (32 Hz and 69 Hz), and c) two levels of duration adjustment for context (50 or 100 msec maximum). The adjustment for context consisted of changing the nominal duration of the stressed vowel depending on the vowel type and following consonant. The figures shown in parentheses are for the highest stress level (1); for lower stress levels (2, 3 and 4) these values are reduced linearly to zero. The adjustment for context is independent of stress level and is used only for stressed vowels. The perturbation of the Fo contour during the stressed portion has the same shape as that shown in Fig. 2.

All possible paired comparisons for each set of 12 utterances were presented to six listeners who were required to choose that item of the pair which sounded more natural (i.e. less machine-like). The subjects were told in advance which word was intended to be emphasized in each of the three test sentences. The results followed a consistent pattern for all three sentences and for 5 of the 6 subjects. There was a strong preference for the smaller Fo increment and a fairly strong preference for the shortest increment in duration. On average, subjects tended to favor the larger context adjustment, but the preferences were not very strong.

The preferred parameter values appear quite reasonable in comparison with other data on synthesis and analysis of stress[5,6,7,8]. Other experimenters[4,7] have noted that changes in Fo of any size are sufficient for denoting stress; the data of this experiment show that for a given stress level the preference is for smaller Fo increments in order to improve naturalness. The preference for a small nominal vowel duration with the larger adjustment for context reflects the balance between these two factors in English. A large nominal duration tends to swamp the context adjustment producing unnaturally long vowels whereas too large a context adjustment would eliminate the common feature of a given stress level - consistent vowel lengthening.

A check on the results of the experiment was carried out in which five new sentences were generated using both the old and new schemes. A similar paired comparison study showed a very strong preference for the new scheme.

Future Directions

Increments in Fo and in duration are only two of the possible factors governing generation of suprasegmental features. Other possible factors include rhythm (i.e. some basic interval between selected segments), the shape of the Fo increment pulse, intensity modulation of the output, and Fo irregularity. Additional refinements in the factors already considered may also be necessary.

Concurrent with this study is an investigation of Fo contours of real speech with selected stress patterns. The purposes of this study are to obtain quantitative data on the variability of the Fo contour across speakers and repetitions by the same speaker and to determine which aspects of the contour are basic to the stress pattern. The results of the study should provide a basis for evaluating the model used in generating the Fo contour as well as indicating which of the above-mentioned factors merit detailed study.

A preliminary analysis of the three sentences used in this investigation has already been carried out using three speakers repeating each utterance three times. A simple yet highly efficient pitch detector was used.[9] The results indicate that the stressed vowel regions of the Fo contours are easily identifiable across speakers and replications. Furthermore, the variability of the contours during voiced, unstressed regions appears to be greater than that during voiced, stressed regions. This suggests that it may be more important to model the Fo contour accurately during the stressed rather than unstressed portion. A typical family of contours for the sentence THIS IS AN OLIVE (with stress on the first syllable of OLIVE) is shown in Fig. 3. The voiced portions have been normalized to the same average time interval for purposes of comparison.

Conclusions

A perceptual experiment was carried out in order to refine the suprasegmental rules used in a procedure for speech synthesis by rule. The parameters considered were increments in fundamental frequency and increases in vowel duration for each of four stress levels. Subsequent experimentation showed that the new rules led to a substantial improvement in the quality of the speech.

Further refinements in the model are clearly necessary and comparisons with suprasegmental features of real speech are currently being made. The results of a preliminary study indicate that the Fo contours of real speech exhibit less variability (across subjects and replications) during stressed regions than during unstressed regions.

References

1. L. R. Rabiner, Speech Synthesis by Rule: An Acoustic Domain Approach, Bell System Tech. J., 47, 17-37, 1968.

2. B. Lindblom, Spectrographic Study of Vowel Reduction, J. Acoust. Soc. Am., 35, 1773-1781, 1963.

3. S. Öhman, Coarticulation in VCV Utterances: Spectrographic Measurements, J. Acoust. Soc. Am., 39, 151-168, 1966.

4. P. Lieberman, Intonation, Perception and Language, MIT Press, 1967.

5. P. Lieberman, Some Acoustic Correlates of Word Stress in American English, J. Acoust. Soc. Am., 32, 451-454, 1960.

6. A. House, On Vowel Duration in English, J. Acoust. Soc. Am., 33, 1174-1178, 1961.

7. D. B. Fry, Experiments in the Perception of Stress, Lang. and Speech 1, 126-152, 1958.

8. I. G. Mattingly, Synthesis by Rule of Prosodic Features, Lang. and Speech 9, 1-13, 1966.

9. B. Gold, Description of a Computer Program for Pitch Detection, Fourth Intl. Cong. on Acoustics, Copenhagen, 1962.

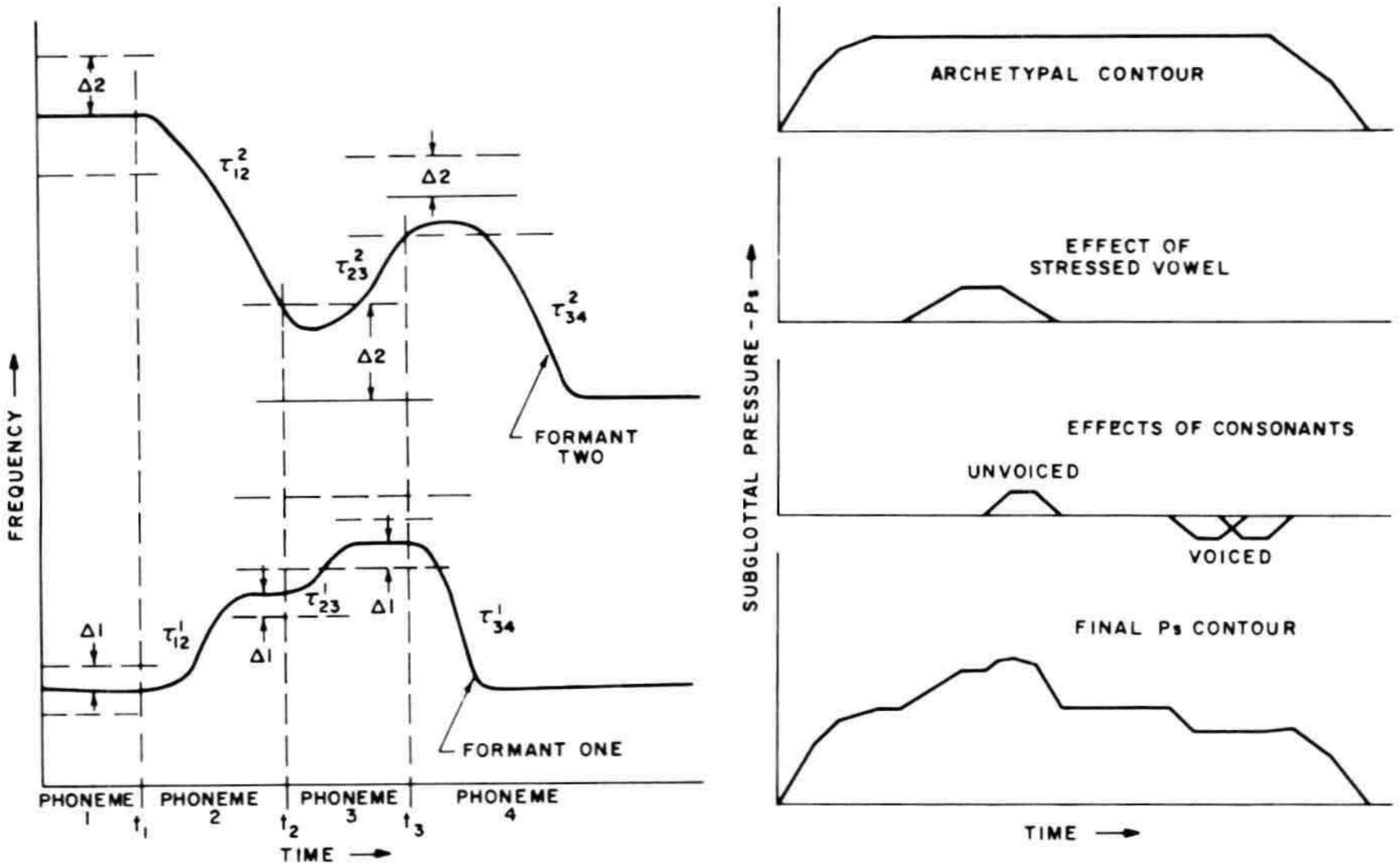


FIG.1 EXAMPLE OF FORMANT CONTOURS

FIG. 2 Ps CONTOUR COMPOSITION

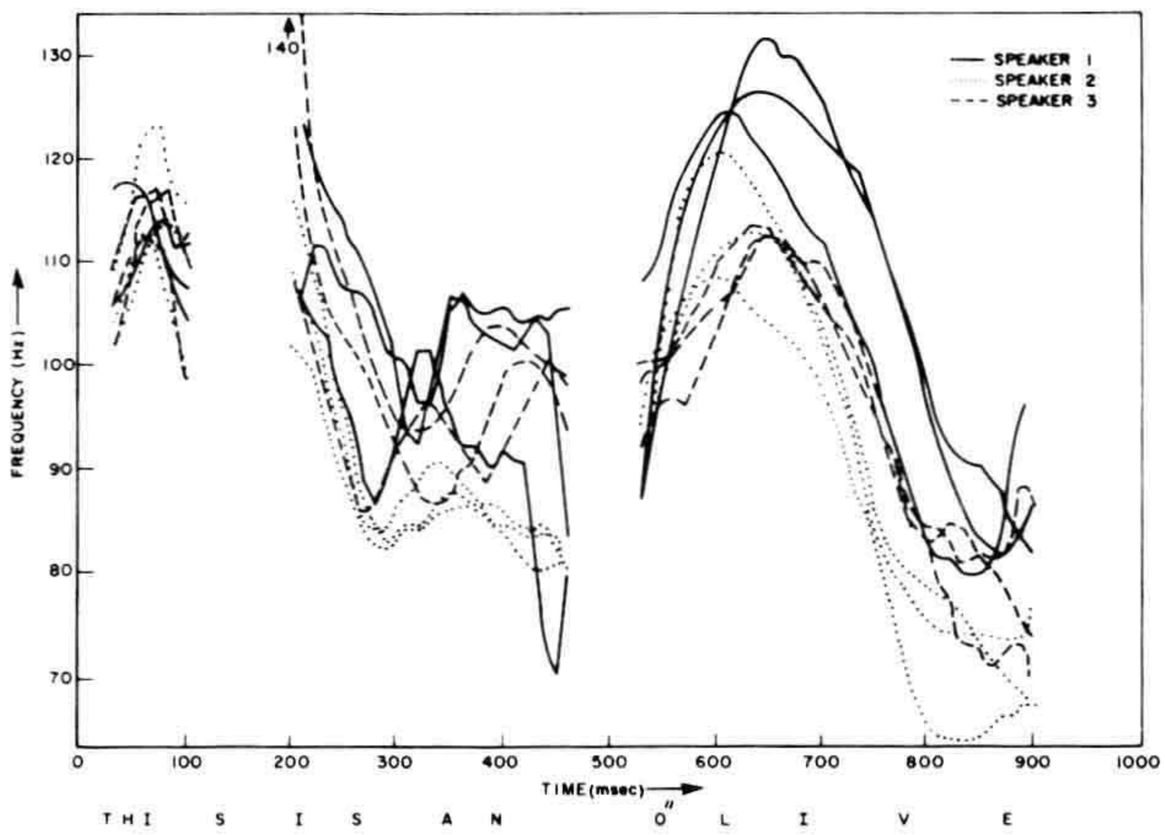


FIG. 3 FAMILY OF FUNDAMENTAL FREQUENCY CONTOURS

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-15

Speech Synthesis Using a Structured Phoneme Classification Scheme

P. V. S. Rao

Fellow

R. B. Thosar

Visiting Member

Tata Institute of Fundamental Research

Bombay, India

Introduction

Most noteworthy among early attempts to systematically synthesize speech are the "brush-painting" experiments at Haskins Laboratories [1]. Synthesis rules here were broad hints for the experimenter to hand-paint phoneme transitions. Similar rules are integrally incorporated into a computer program in the Holmes-Mattingley-Shearme system [2] which, given values for a large number of relevant parameters for each phoneme, comes out with a spectrographic specification for synthesized speech, to be fed into an analog play-back unit. The "diphonic" approach [3], while sidestepping the concatenation problem, suffers from other wellknown serious limitations.

This paper outlines a synthesis scheme which has been evolved using a flexible programming package called SPEECH (for Speech Playback Evaluation and Experimental Concatenation Handler), developed specially for the purpose [4]. The main features of this approach are: (1) phonemes are divided into equivalence classes and subclasses to facilitate formulation and application of context dependency rules; (2) these rules are explicitly provided for as separate subroutines. Further synthesis, even upto the level of the final acoustic signal, is

accomplished within the computer. Only the final digital to analog conversion is done by hardware, thus minimising equipment imposed limitations to quality and flexibility. The input message string can be typed in from the console typewriter of the computer.

The Synthesis Scheme

The main problem in speech synthesis is the faithful replication of neighbourhood effects that occur in human speech, either at the neuromotor level, due to partial overlapping of motor commands pertaining to adjacent phonemes in connected speech as issued by the brain (anticipatory articulation), or at the articulatory level, due to the finite time constants of movement of the articulatory organs. Overlapping at the neuromotor level should be more pronounced for commands pertaining to different articulatory organs. It would thus seem logical to assume that phonemes with pronounced similarities in their articulatory or neuromotor representations have similar combinatory effects on adjacent phonemes. The present effort, therefore, has been to make a structured classification of phonemes into equivalence classes and subclasses mainly to facilitate the formulation of synthesis rules (Fig. 1). The subgroup "vowels", for instance, consists of phonemes which, in conjunction with other phonemes, exhibit continuous (and linear) transitions of formant

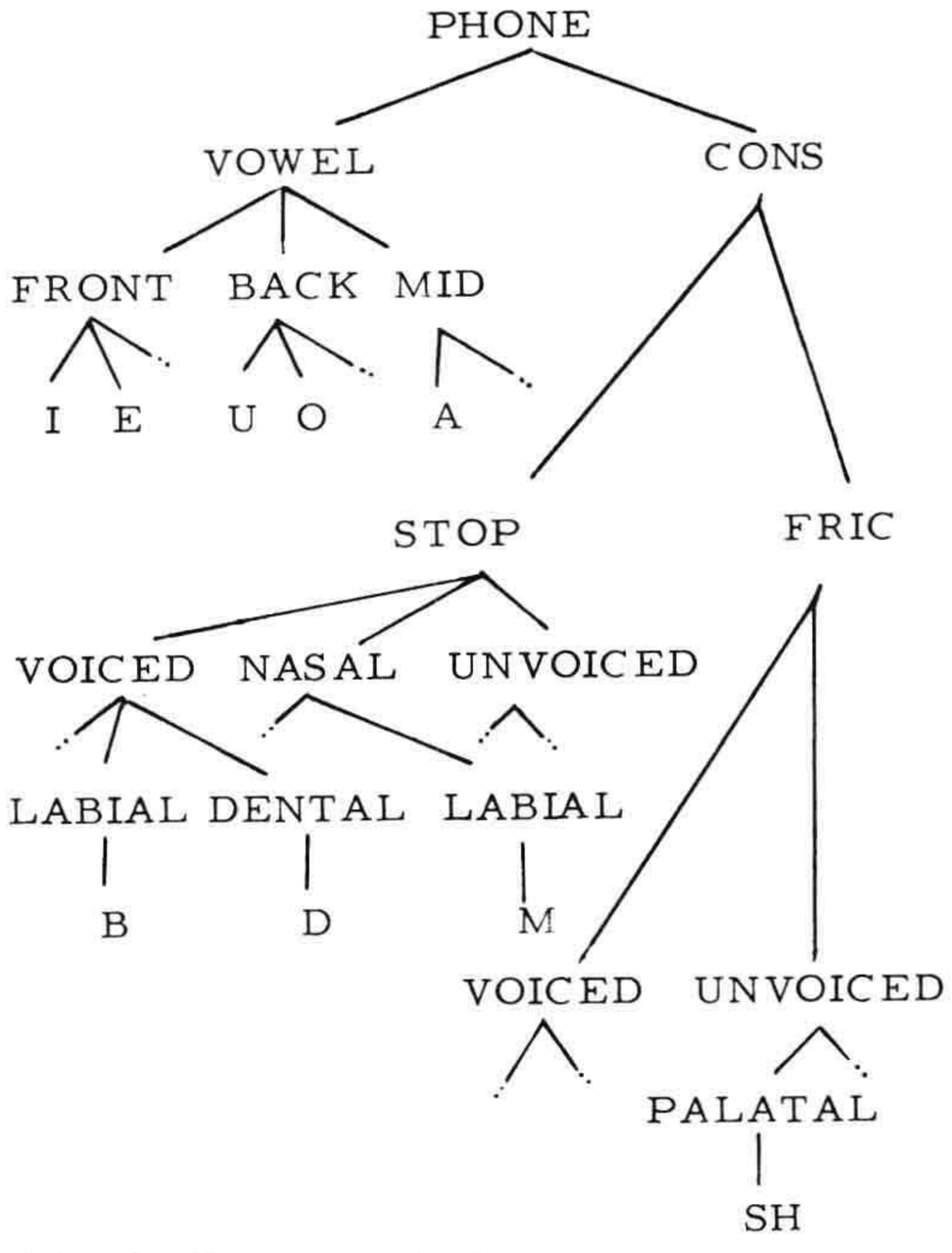


Fig. 1. Structured classification of phonemes (partially shown).

frequencies and amplitudes. The stop-fricative classification for consonants is on the basis of similar first formant transitions and the labial-dental-velar classification on the basis of second formant transitions. That such a classification is possible and runs generally parallel to those based on articulatory and acoustic criteria demonstrates the validity of this approach.

Steady state characteristics (such as formant frequencies) for vowels and context invariant properties (such as plosive burst frequencies) for consonants are specified for each phoneme, while concatenation parameters (such as formant loci and transition times) are specified for each group and subgroup. Concatenation can be effected using this information, by formulating well-defined rules for the purpose.

Concatenation Procedure

Vowel-consonant concatenation is used to illustrate this procedure (Fig. 2). Initial or terminal frequencies for CV or VC formant transitions are obtained from loci (L) specified in common for each subgroup (first formant depending on manner and second formant on place of articulation), by specifying a transition fraction(F) as a concatenation parameter. Formant loci can thus be common for several vowels. Subdivisions of vowels (as frontal, medial and back) allows for multivalued loci as in the case of palatals.

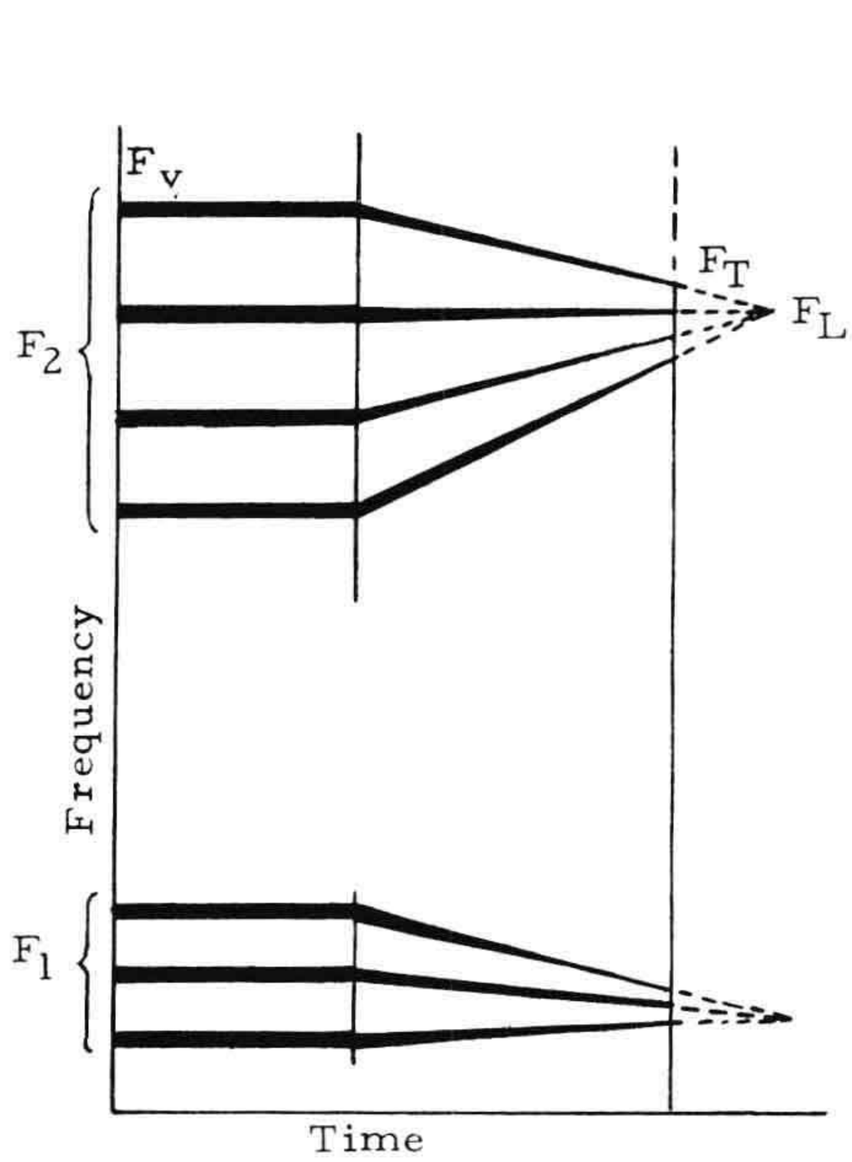


Fig. 2. Computation of formant transitions

$F_T = F_v - F(F_v - F_L)$

Formant transition times can be different for fricatives and stops. The formant amplitude contour jumps discontinuously to zero for stops (and nasals) at the phoneme-boundary, while it reduces linearly for fricatives.

Prosodic Features

Three types of pitch modulation are to be incorporated into the system: a continuous fall for declarative sentences, a rise in the final syllable for interrogative sentences, and a provision for stressing individual syllables to any desired extent by raised pitch.

Conclusion

Individual words, nonsense syllables and connected sentences have been synthesised using this scheme, and are intelligible and clear. Introduction of prosodic effects by pitch modulation has been found to add substantially to the intelligibility and naturalness of synthesised speech.

The ease with which modifications can be made to the classification scheme as well as to the rules make the system extremely flexible. It is hoped that such a scheme will give more insight into the speech recognition problem as well.

1. Liberman, A. et al., Minimal Rules for Synthesizing Speech, J. A. S. A., 1959.

2. Holmes, J. N. et al., Speech Synthesis by Rule, Language and Speech, 1964.

3. Estes, S. E. et al., Speech Synthesis from Stored Data, I. B. M. J. of Research and Development, 1964.

4. Rao, P. V. S. and Thosar, R. B., SPEECH: A Software Tool for Speech Synthesis Experiments. Technical Report No. 38, Computer Group, Tata Institute of Fundamental Research, Bombay 5.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-16

Multiplexing Speech Synthesizer

Kazuo Nakata and Tanetoshi Miura

Central Research Laboratory, Hitachi, Ltd.

Introduction

Two points should be considered seriously in the practical application of the speech synthesizer as an audio output unit of an information processing system. One is, needless to say, the quality of synthesized speech and the other is multiplexing operation.

Multiplexing operation will be realized by two modes: multiplexing synthesis and multiplexing control. The former needs only one synthesizer but the latter needs a number of them.

Usually, a compilation of pre-recorded units of speech is easily operated in multiplexing synthesis mode but a speech synthesizer can work only in multiplexing control mode.

Recently, however, new ideas of multiplexing speech synthesis have been proposed in Japan, and this paper describes one of these new techniques of multiplexing speech synthesis.

Compilation of Acoustical Elements of Speech

The key point of this new method is the realization of speech synthesis in acoustical level by the use of compilation technique and the control of compilation process by pitch.

Our proposed method, therefore, may be called "pitch-controlled compilation of

pre-recorded acoustical elements of speech".

Acoustical elements of speech mean a series of damped sinusoids of various frequencies and band-passed noises of various pass-frequencies, and they are recorded on a memory drum as shown in Fig. 1.

Proper compilation of these acoustical elements, as shown in Fig. 2, must reproduce speech-like sounds.

Then the first half of the method - compilation of acoustical elements of speech - will be realized, but the second half - pitch-controlled compilation - will not be realized by this simple technique.

Pitch-Controlled Compilation

The principle of pitch-controlled compilation is shown in Fig. 3. N readout heads are equally spaced around the memory drum on which a damped sinusoid of duration T sec. is pre-recorded.

If the first head is always assigned to read out the signal, the output is the damped sinusoid repeated with a period of T sec. and this means constant pitch compilation. If relative position of readout head is suitably controlled by pitch information, pitch synchronous readout of a damped sinusoid is approximately realized.

The pitch of compiled speech thus controlled can change only in quantized fashion with a minimal step of $\tau = T / N$ sec. rather than continuously, but this method does not need any extra buffer memory for pitch control.

Practically, an equivalent operation can be performed by the use of time division multiplexing recording of samples and this modification needs only one readout head.

Computer Simulation and Hearing Tests

First of all, we must examine how good is the quality of synthesized speech by the method. Evaluation of quality is carried out with synthesized samples produced by computer simulation of the method.

The conditions of computer simulation are shown in Table 1.

The results of hearing tests convince us that the quality of the speech is

acceptable from the practival viewpoint.

A sonagram of synthesized sample is shown in Fig. 4.

Concluding Remarks

A series of experimental studies and hearing tests led us to the following conclusions.

1) The proposed method is worthy of further study as a multiplexing speech output unit in practical application.

2) Relative amplitude control of each formant component might be simplified by a simple de-emphasis after analog summation of three formant components.

3) The appropriate steps of time interval of pitch-controlled compilation are about 32 equal divisions of 10 msec. each.

4) Overlapping of residual wave form in pitch controlled compilation does not show any noticeable improvement in speech quality compared with simple clear-and-add compilation.

All these experimental findings combine to improve the method, simplifying it and making it worth considering for practical application.

Table 1. Conditions of Computer Simulation.

Elements	Number of Elements	Range and Step
First Formant	16	200~950 Hz, f= 50 Hz
Second Formant	16	800~2200 Hz, $\Delta f/f$= .07
Third Formant	8	2200~3600 Hz, $\Delta f/f$= .07
Noise	8*	(*16 will be sufficient)

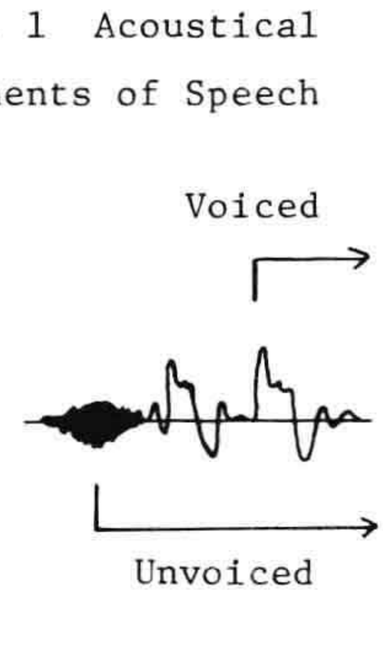


Fig. 1 Acoustical elements of Speech

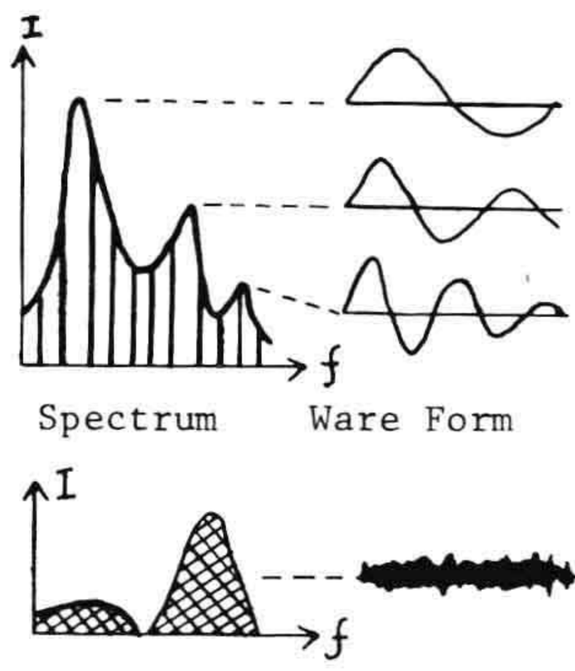
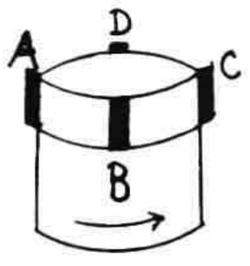
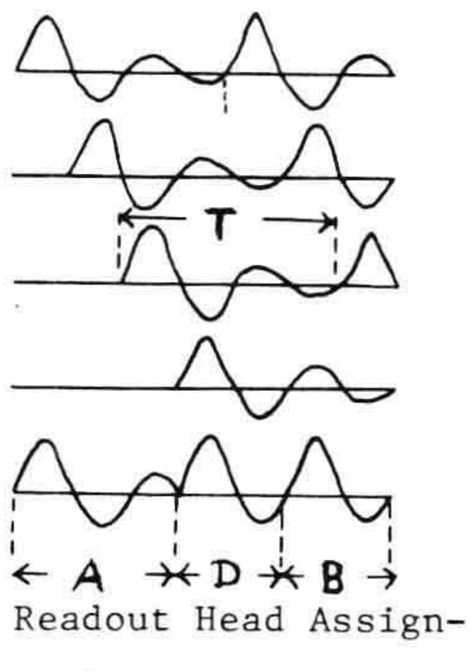


Fig. 3 Pitch-Controlled Compilation





Readout Head Assignment

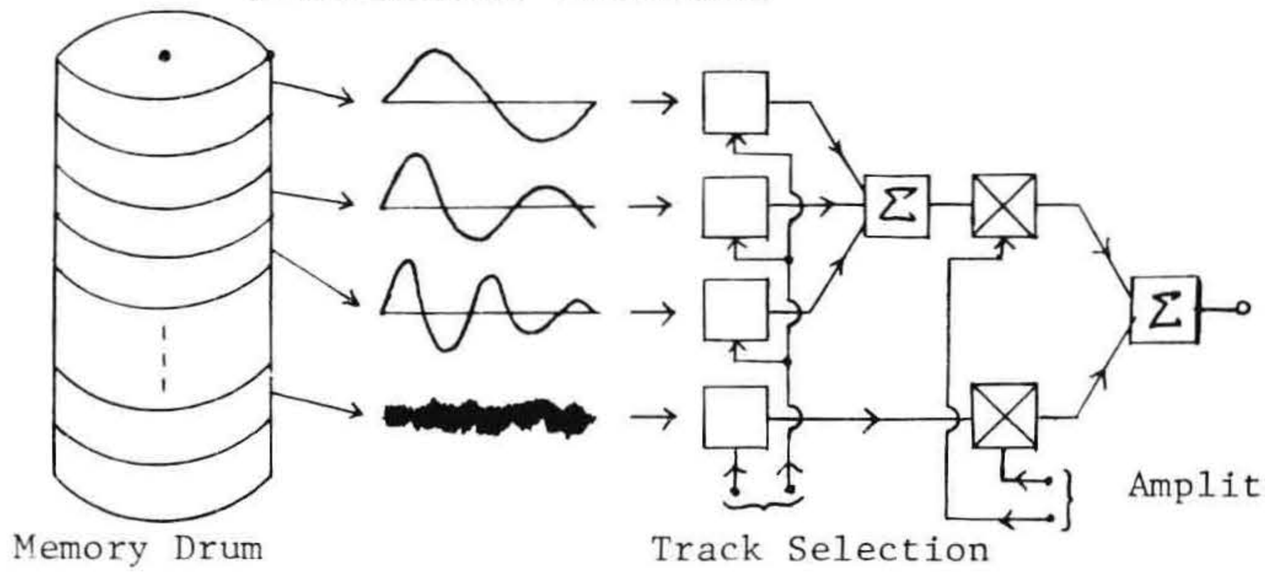


Fig. 2 Speech Synthesis by Compilation of Acoustical elements

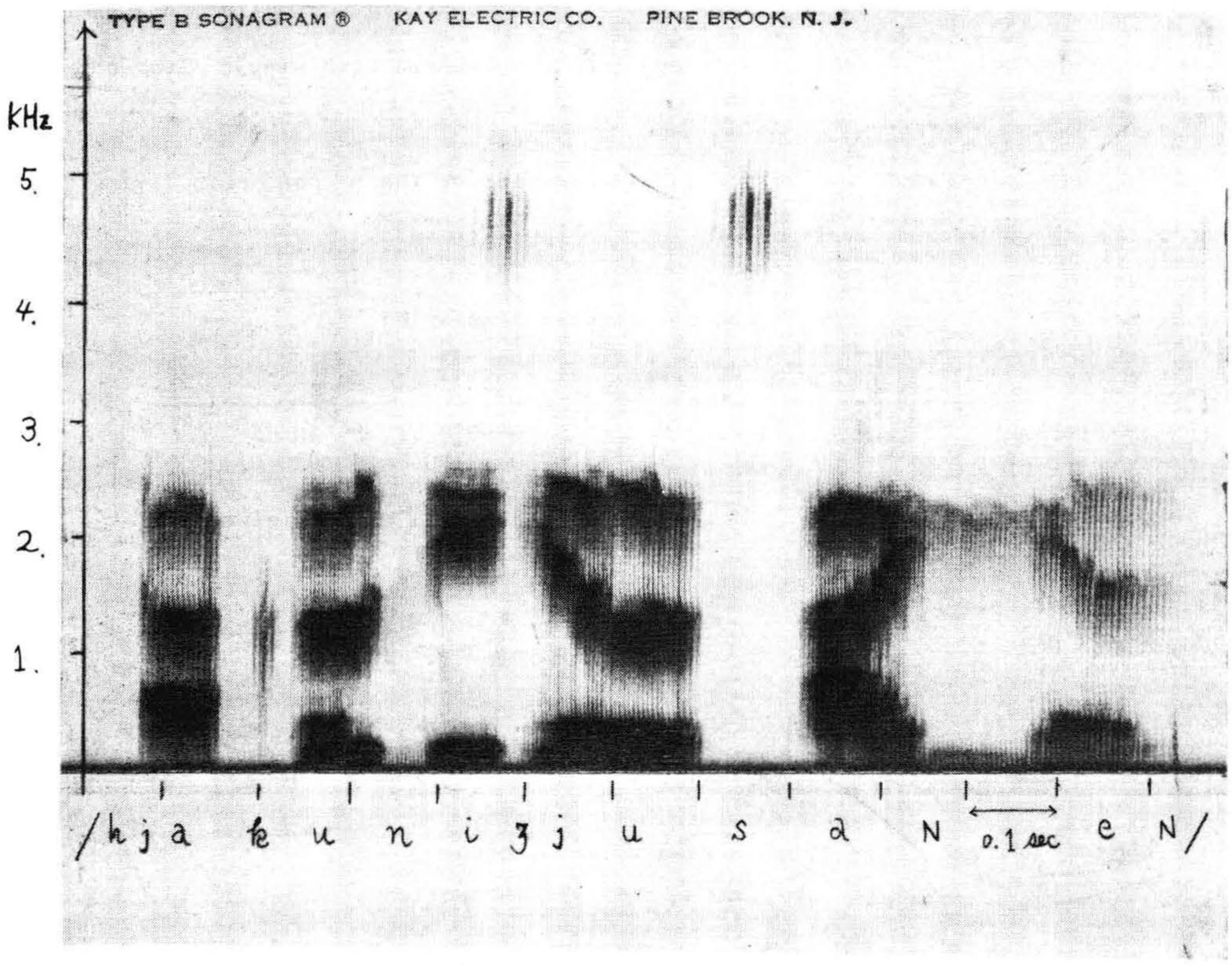


Fig. 4 A Spectrogram of Synthetic Speech.

(Japanese / hjaku nizjusanen /)

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

The Use of Teaching Machines for Training Experiments in Speech Perception

Karl-Erik Spens and Arne Risberg

Research associates of the Department of Speech Communication, Royal Institute of Technology (KTH), Stockholm, Sweden

Introduction

There is at present a great deal of interest in developing new technical aids for the deaf and hard of hearing in which the speech signal is analyzed and some important elements are extracted and then transmitted to the hearing impaired subject as a new acoustic, tactual, or visual signal. In the development and evaluation of these aids it is necessary to run a large number of TTT-experiments.* It is also necessary to define these experiments very carefully so that different solutions can be easily compared. These considerations and also the need of reducing the amount of man power needed to run the experiments have resulted in the development of a number of TTT-machines at the Department of Speech Communication in Stockholm. In this paper will two of them be described.

TTT-machine for individual experimentation

A block diagram is shown in Fig. 1. The tape-recorder has got two tracks, one for the stimuli and the other for the control code. 15 answering alternatives were decided to be enough which meant that a 4-bit binary control code could be used. The control code is a 4-frequency tone code which starts with a 0.4 sec segment for the control logic followed by a 0.4 sec segment which gives information about the correct answer. The control code always precedes the stimulus.

* TTT stands for Teaching, Training, and Testing.

The tape-recorder automatically stops when it reaches the control code for the next stimulus which means that the TTT procedure is self-paced. The subject answers on a set of 15 pushbuttons. The answer is compared with the stored code and the result is fed to counters which indicate the number of answers and the number of correct answers and also to a program selector. Three programs can be used.

1. Prompting program (teaching). The subject starts the tape-recorder and is immediately informed by the corresponding lamp about the correct answer before the stimulus is heard. The subject must answer by using the corresponding button and can then proceed to the next stimulus by means of the "NEXT" button. Before giving answers the stimulus can be repeated by means of the "REPEAT" button.

2. Confirmation program (training). The subject hears the stimulus and answers on the pushbuttons. He then receives the correct answer and if his answer was correct he can proceed to the next stimulus by means of the "NEXT" button. If his answer was wrong he has to repeat the stimulus by means of the "REPEAT" button and then give a correct answer before he can proceed to the next stimulus. He can also use the repeating function before giving any answer.

3. Test program (testing). The subject hears the stimulus and gives his answer and then proceeds to the next stimulus by means of the "NEXT" button. Repetition is not possible.

TTT-machine for group experimentation

Block diagram of the machine is shown in Fig. 2. This TTT-machine is meant for groups with maximally nine subjects. Each subject has his own answering box where there are pushbuttons corresponding to the answer alternatives, which, in this case, are presented on a screen by means of the projector. As in the other TTT-machine the tape-recorder has two tracks. The tape-recorder and the way of programming the tape with stimulus and control codes are principally the same as in the other machine.

The answering alternatives are changed with each stimulus, therefore the number of alternatives in the whole experiment is not limited as in experiments with the other TTT-machine. At present there are three pushbuttons on every answering

box. If needed, this can easily be extended to maximally 16. This TTT-machine is controlled by the teacher, that is, the teacher pushes the "NEXT STIMULUS" button. He also registers the answers from each subject. These answers are presented on a lamp matrix by marking the place of every lit-up lamp on a paper placed over the lamp matrix. These registrations can then be read by an optical reader and handled by a computer. Information about the answers being correct or not is given the subject by a green and a red lamp on the answering box. These two lamps are controlled by the output signal from the comparison between the subject's answer and the stored control code.

The training sequence, recently used at the Manilla-School for hard of hearing children in Stockholm in experiments referred by A. Risberg at this congress, was as follows:

1. The teacher pushes the "NEXT STIMULUS" button.
2. A picture with two words which only differs in the vowel or in a voiced or unvoiced sound is presented on the screen. The tape-recorder starts and the subjects hear one of the words. The control code is stored and the key lamp corresponding to the correct answer lights up. The teacher puts a sheet of paper over the lamp matrix.
3. The subjects answer by pushing the button corresponding to the word they think they have heard. Is the answer correct the green lamp in the answering box will light up and if it is wrong the red one will light up. The answers are also registered on the lamp matrix.
4. The teacher marks the lit lamps. Then he pushes the "NEXT STIMULUS" button.
5. The picture changes and now only the correct word is presented on the screen and repeated by the tape-recorder.
6. The sequence starts again.

One of the greatest advantages with this machine is the modest amount of manpower needed to collect and handle the great amount of registrations needed to get significant results from experiments of this type.

The kind of competition which automatically exists between TTT-machines and

the subjects seems to be very stimulating to the subjects and keeps their motivation on a high level. The hard of hearing children at the Manilla-School have in spite of the intensity of the recent training period had an extremely positive attitude to the TTT-machines.

Acknowledgments

The development of the TTT-machines has been supported by the Swedish Council of Applied Research and it has been done in cooperation with the Dept. of Education at the Uppsala University. The recent TTT-experiments have been done at the school for hard of hearing in Stockholm (Manilla-School). We are very grateful for the positive attitude from everyone at the school and especially for the help we have got from the teacher of the deaf, Mr. C.B. Flodell.

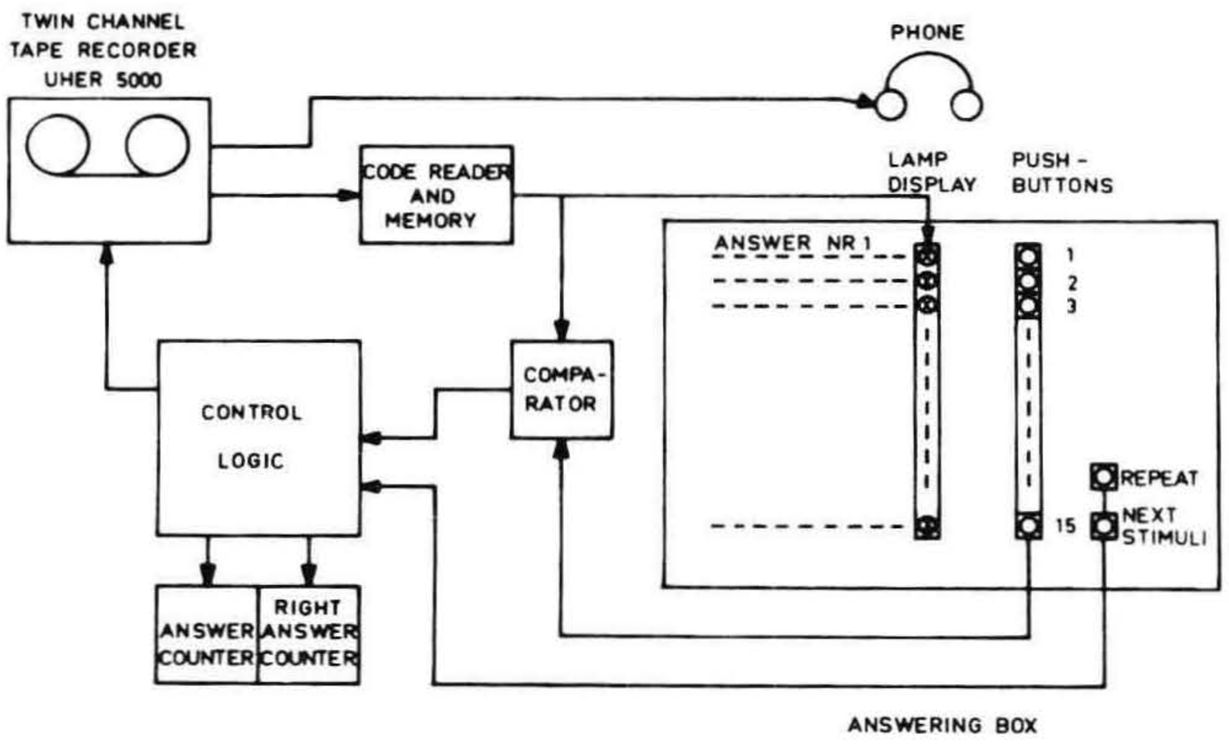


Fig. 1

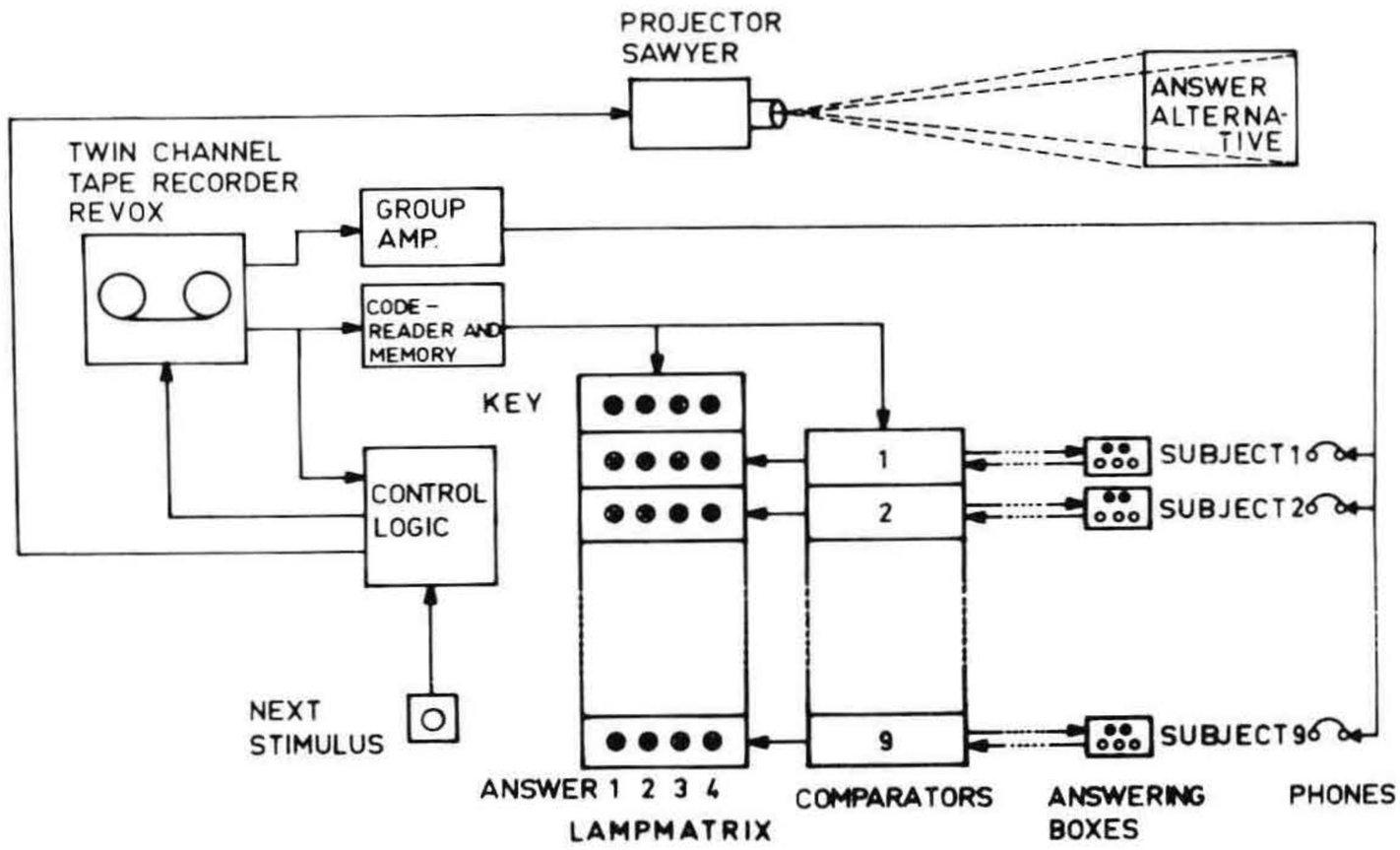


Fig. 2

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

B-5-18

Periodic-nonperiodic test of hearing capacity

Arne Risberg

Research associate of the Department of

Speech Communication, Royal Institute of

Technology, Stockholm, Sweden

The term "hearing capacity" has been used by Pimonow (1965) to describe the possibility of the auditory system to transmit information in the same was as "channel capacity" is used in telecommunication. By measuring the time-, frequency-, and amplitude-resolution of an impaired ear it is possible to estimate the number of bits per second that can be transmitted. This information is then used to select an optimum coding strategy in a speech-analyzing hearing aid, that is, a hearing aid where the speech signal is analyzed, some important information extracted and then coded in another form or in a different frequency region.

For the same purpose Pickett (1965) measured the spectrum resolution by hearing impaired subjects for low-pass filtered noise signals. This measurement might be of value for the estimation of the subject's success in using a fricative-transposing hearing aid (Johansson 1961, Pimonow 1965).

The purpose of the measurements reported (PNP-test) here is the same as the above-mentioned measurement but it is also hoped that the test described can be used to establish if a subject has residual hearing or if the response in the tone audiogram is due to sensation of vibration.

Test procedure and results

A block diagram of the test equipment is shown in Fig. 1. The rise and decay times of the gate are 2 msec. The test starts with the establishment of the thres-

hold for the periodic signal by means of the attenuator "C". Signal duration 1000 msec and periodicity usually 200 pulses /sec. The next step is to establish the threshold for the nonperiodic signal (noise) by means of attenuator "A" and the same duration. The test level is set 15 dB above the thresholds by means of the attenuator "C".

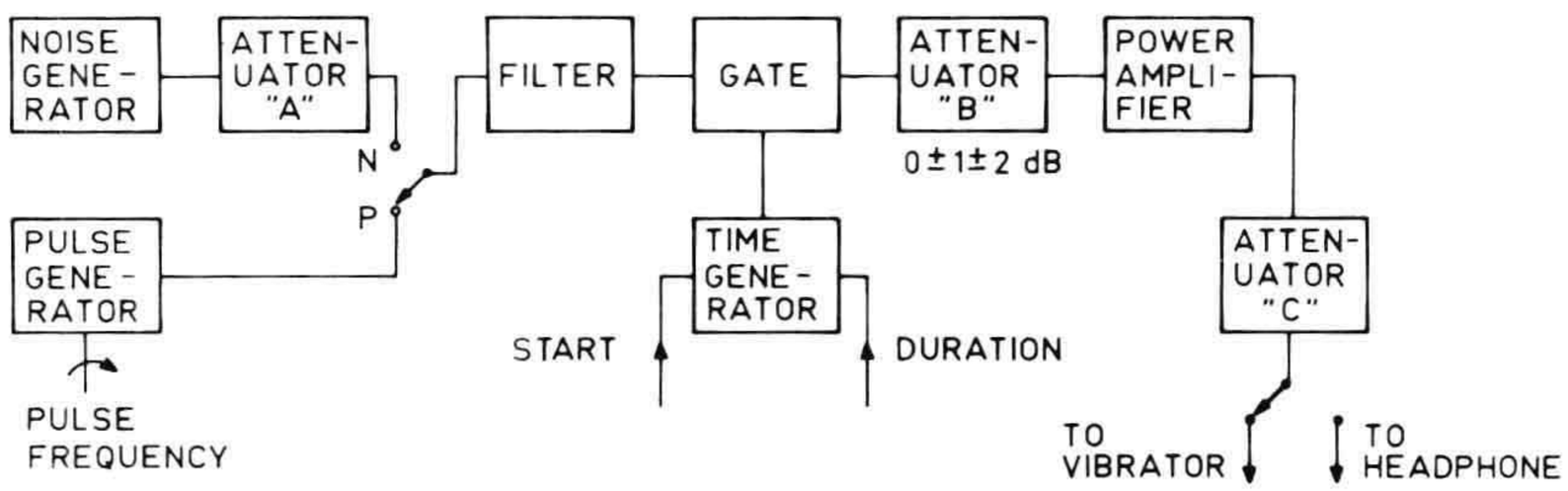


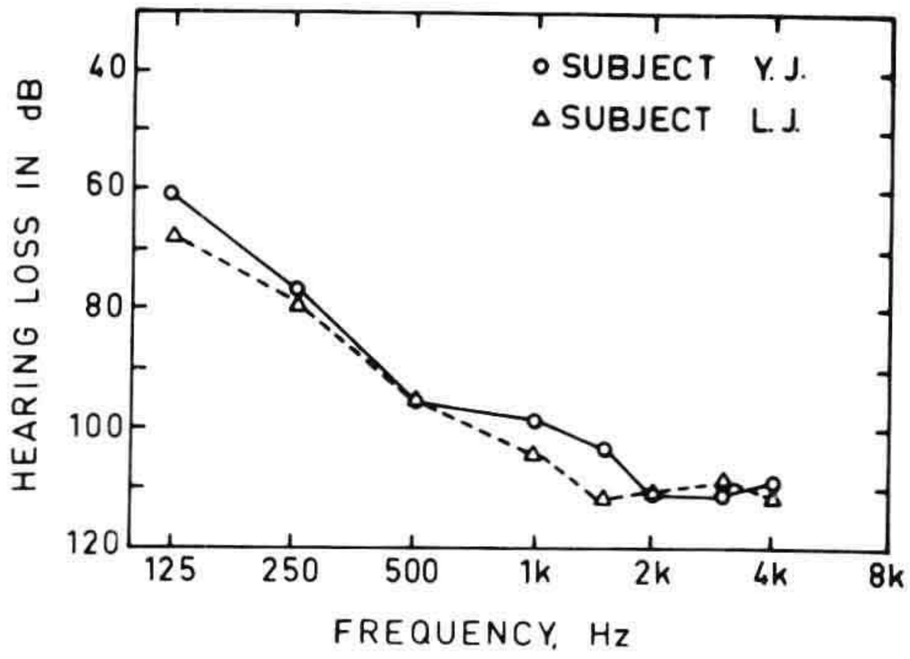
Fig. 1. Block diagram of test equipment.

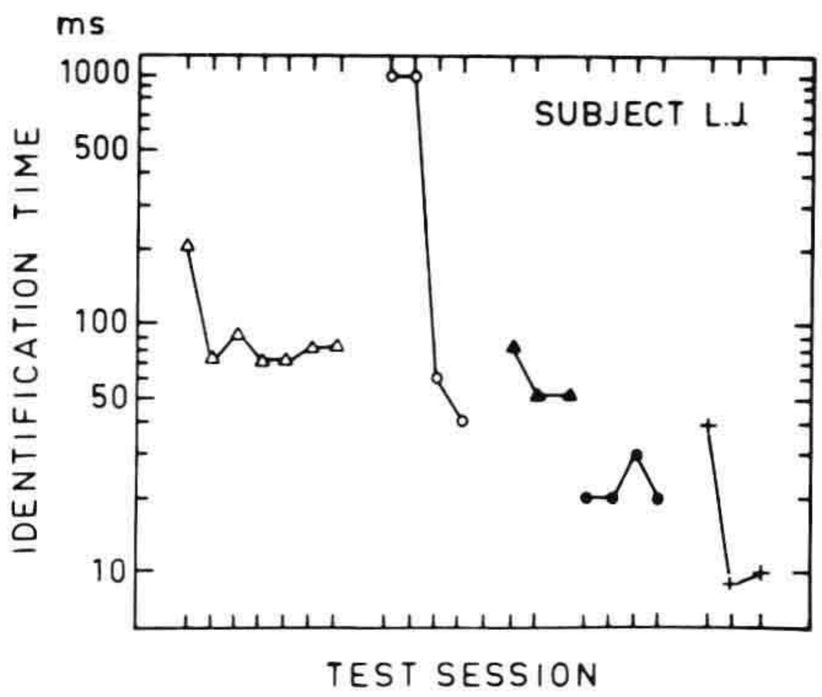
A block of ten stimuli is presented to the subject with random variation of periodic and nonperiodic signal and also random variation on attenuator "B". The starting duration is 1000 msec and the subject answers on two pushbuttons and is immediately informed if he is correct or not. The test continues until the duration T_i is found where he cannot distinguish the two signals with at least 80 % accuracy. The attenuator "B" is used to make it impossible for the subject to use loudness as a clue in the identification.

The duration T_i is measured both for tactual stimulation on a finger tip by means of a vibrator (Goodman V 47) and for stimulation in the ear by headphones (Telphonic TDH 39). In the last measurement a broad-band signal can be used, 100-4000 Hz, or the measurement can be made in two bands 300-800 Hz and 2400-3400 Hz.

In Fig. 2 results from measurements on two deaf-born children, age 13 years, are shown. The high-frequency test could not be made on subject Y.J. due to discomfort and for subject L.J. the test level was 10 dB above the established threshold. All other measurements are made 15 dB above the threshold. For normal

hearing subjects the tactual value will vary between 40 and 80 msec and the auditory values will vary between 10 and 20 msec for the low-frequency band and around 10 msec for the high-frequency band.





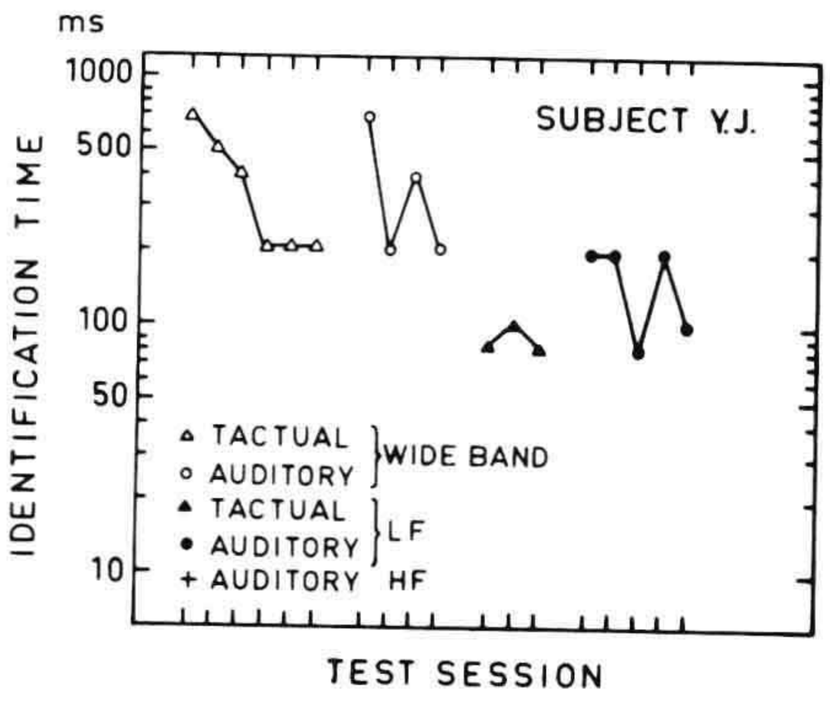


Fig. 2. Pure tone audiogram and the results from the PNP-test for two severely hard of hearing subjects.

The audiograms of the two subjects in Fig. 2 are almost identical or slightly better for subject Y.J. In spite of this great differences are obtained on the PNP-test. Subject L.J. obtains shorter T_i for auditory signals than for tactual both for low and high frequencies and this seems to warrant the conclusion that he has residual hearing for these frequencies. This does not seem to be the case for subject Y.J. The PNP-test then seems to be able to give valuable supplementary information to the pure-tone audiogram.

Correlation with speech perception

To see if the test could be used to predict result in training with a fricative-transposing hearing aid a training experiment was made with a group of six children, age 14 years. The training material consisted of rhyming word-pairs containing the phonemes /l, v - s/. After fricative-transposition it is possible to

distinguish between the two words if the subject can distinguish between a low-frequency periodic signal and a noise signal with the length of a phoneme. The test signals were low-pass filtered 1000 Hz to restrict the information to the low-frequency end. The training was achieved by means of the group teaching machine described in a paper by Spens and Risberg at this congress. The result after four training sessions (15 min each) and the T_i values is shown in the table below.

Subject	Transp. test % correct	Tactual msec	Low freq. msec	High freq. msec
V.J.	83	40	20	10
R.A.	100	40	20	20
T.H.	83	40	20	20
E.K.	50	60	60	-
A.W.	48	40	30	-
S.A.	60	30	30	-

50 % is equal to guessing. For the last three subjects it was not possible to measure the high-frequency values. From the table it seems that it is possible to use the PNP-test to predict result from training with fricative-transposing hearing aids which is of importance if the test can be run on very young children without language. Testsof this type are planned.

Acknowledgments

The research reported has been supported by the Swedish Council of Applied Research. All measurements have been made at the School for the hard of hearing (the Manilla-School) in Stockholm by Miss E. Agelfors. Dipl.ing. J. Mártony has taken a very active part in the development of the test method.

References

Johansson, B.: "A New Coding Amplifier System for the Severely Hard of Hearing", Proc. of the 3rd ICA, Stuttgart 1959 (Amsterdam 1961), II, pp. 655-657.

Pickett, J.M., Daly, R.L., and Brand, S.L.: "Discrimination of Spectral Cut-Off Frequency in Residual Hearing and in Normal Hearing", paper B 33 presented at the 5th ICA, Liège 1965.

Pimonow, L.: "La parole synthétique et son application dans la correction auditive", Annales des Télécommunications 20 (1965), pp. 151-171, No.7-8.

C

SPEECH

CONTENTS

C–5

C–5–1 Test "CIF" standard pour essais d'intelligibilité en espagnol
Carmen Delgado de Guzmán C–1

C–5–2 Influence de la distorsion harmonique d'un système de transmission sur l'intelligibilité
Jaime Pfretzschner, Adelaida Jover ..... –5

P–0–26 The twin-com system
J. J. Geluk ..... –9

C–5–4 Predictive coding of speech signals
Bishnu S. Atal, Manfred R. Schroeder ..... –13

C–5–5 An analysis-synthesis telephony based on maximum likelihood method
Fumitada Itakura, Shuzo Saito ..... –17

C–5–6 La synthese de la parole a partir de digrammes phonetiques
Emile Leipp, M. Castellengo, J. S. Lienard ..... –21

C–5–7 ZEX coding for speech transmission
Shuzo Saito, Kiyoshi Hashimoto ..... –25

C–5–8 Influence of bandwidth and reverberation time on preference evaluation of speech
James E. West, Gerhard M. Sessler ..... –29

C-5-9 Recherches récentes dans le domaine de l'Acoustique et du Traitement des signaux de parole menées au C. N. E. T.
Gilbert Ferrieu C-33

C-5-10 Universal phonetics as an explanatory theory
Victoria A. Fromkin ..... -37

C-5-11 Speech wave, articulation and their linguistic function
Józef Wierzchowski, Bozena Wierzchowska ..... -41

C-5-12 On learning linguistic concepts
Ernest Colton ..... -45

C-5-13 System-theoretical considerations on speech generation, transmission and processing
Syuji Inomata ..... -49

C-5-14 Phonèmes et phonétographe
Jean Albert Dreyfus-graf ..... -53

C-5-15 Consonantal clusters and connected speech recognition
D. Raj Reddy ..... -57

C-5-17 The optimum filter for frequency analysis of speech
Yoshinori Isomichi ..... -61

C-5-18 Real time colored display for sound spectrograms
Tsuneji Koshikawa, Yasuaki Nakano, Takeshi Nakayama, Tanetoshi Miura ..... -65

C-5-19 Analysis of speech sounds in ultrasonic frequency range
Helen N. Miasnikova, Lev L. Miasnikov, Nicolai P. Gromov ..... -69

P-0-35 Acoustical problems in non-verbal communication
H. Burris-Meyer ..... -73

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

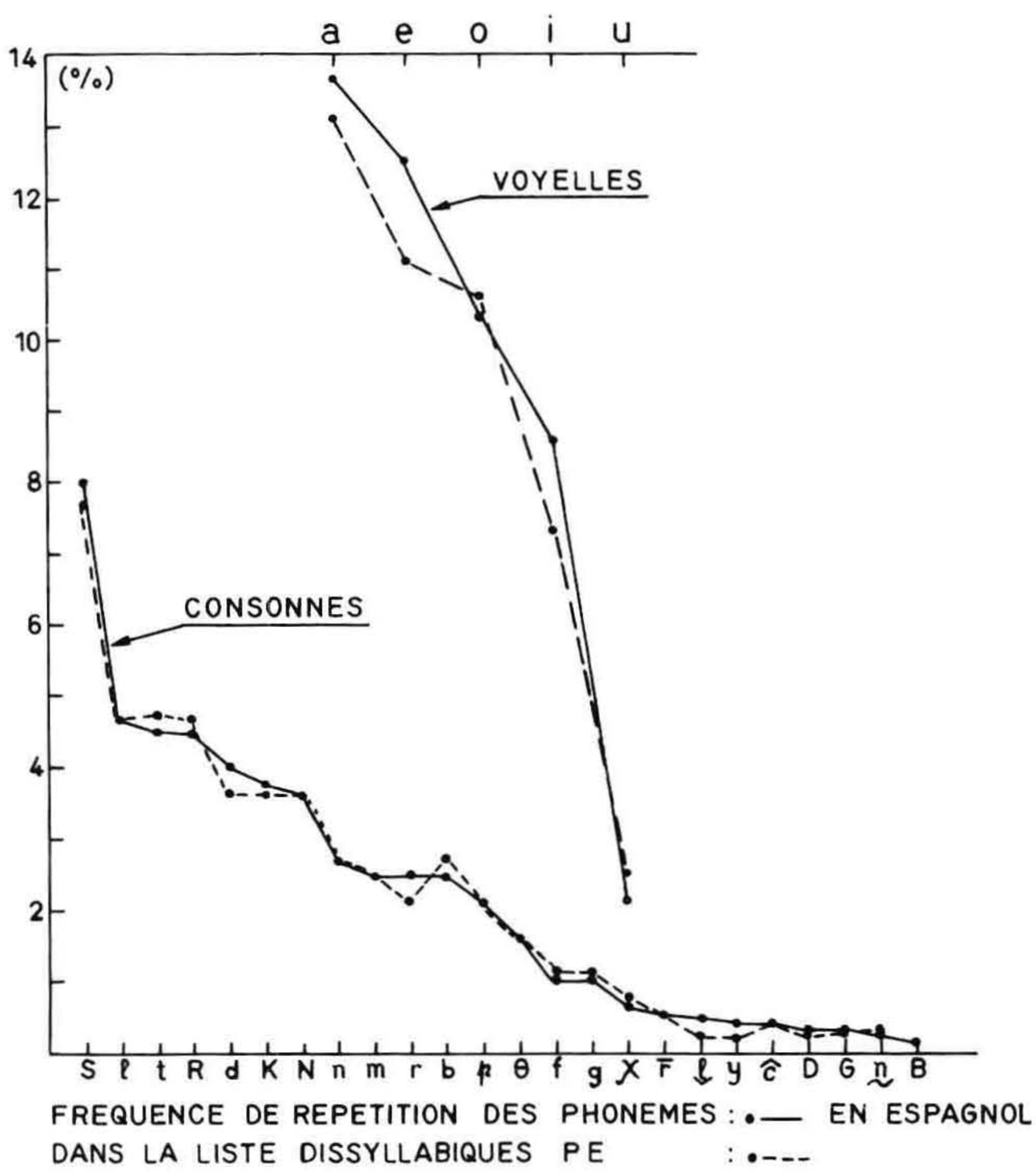
C-5-1

Test CIF standard pour essais d'intelligibilité en espagnol

Delgado, Carmen

Département d'Acoustique
Centro de Investigaciones Fisicas "L.Torres Quevedo"
Madrid.- Espagne

1.-Test CIF.- Partant de la fréquence de répetition des phonèmes dans la langue espagnole, et des mesures subjectives, on a determiné deux listes de mots, monosyllabiques et dissyllabiques pour épreuves d'intelligibilité.



FREQUENCE DE REPETITION DES PHONEMES : •— EN ESPAGNOL
DANS LA LISTE DISSYLLABIQUES PE : •---

FIG. 1

Le test des dissyllabiques se compose de 100 mots distribués en groupes de 10, avec lequel on est arrivé a une distribution de phonèmes - très proche de celle qui existe dans l'espagnol écrit; on peut donc le considérer comme phonétiquement equilibré(fg.1).

La liste correspondant aux monosyllabes n'est composée que de 35 mots, puisqu'en espagnol ce genre de mots n'est pas fréquent. A cause du nombre si réduit et de la structure phonétique propre de chaque mot, on ne peut pas arriver à un equilibre phonétique.

TEST CIF

Monosyllabiques

M1	M2	M3	M4	M5	M6	M7
sal	tres	res	flan	crin	son	par
tren	mal	tul	ser	mar	red	ron
vid	fin	can	tos	sed	bar	mes
pez	pan	te	vil	col	gas	gris
flor	dos	paz	mas	sur	tez	sol

Dissyllabiques PE

B1	B2	B3	B4	B5
tomo	grifo	losa	verde	isla
lumbre	senda	cima	silla	luna
casco	llave	hambre	lanza	peso
mero	perro	duque	tasa	dique
coche	queso	pino	pena	sala
friso	mosca	fresa	denso	nido
vista	rima	nave	rosca	eter
dardo	yunque	cara	torre	vaso
lana	frente	silo	guiso	cero
iris	tino	mente	parche	trampa

B6	B7	B8	B9	B10
sino	muro	tanque	seda	tira
lejos	gente	piso	niño	sombra
dulce	raso	cumbre	tumba	limbo
seta	lino	vela	cana	toga
calle	termo	lista	lima	dedo
mesa	debil	trece	rosa	hijo
tiza	prado	sano	clero	ganso
aro	tecla	util	eje	cera
ritmo	nudo	cabo	disco	pluma
soga	fila	pera	linde	signo

2.- Essais

Les essais d'intelligibilité ont été réalisés dans la Cabine d'Essais Subjectifs du CIF "L. Torres Quevedo", en utilisant des écouteurs et en employant le test enregistré d'avance sur bande magnétique.

L' auditoire se composait de 15 personnes otologiquement normales, âgées entre 20 et 30 ans, appartenant aux differents laboratoires du Centre.

On donne les résultats des essais réalisés avec les listes de mots mentionnées,ci-dessus, ainsi que des essais réalisés avec une liste de 35 dissyllabes choisis au hasard.

Dans la fig.2a on présente les courbes du taux d'intelligibilité obtenu pour tout le rang des fréquences audibles, en fonction du niveau de la pression sonore, qui a varié de 5 en 5 dB jusqu'à 60 dB.

On observe une grande coincidence entre les courbes obtenues avec les mots monosyllabiques et avec l'ensemble des dissyllabes choisis au

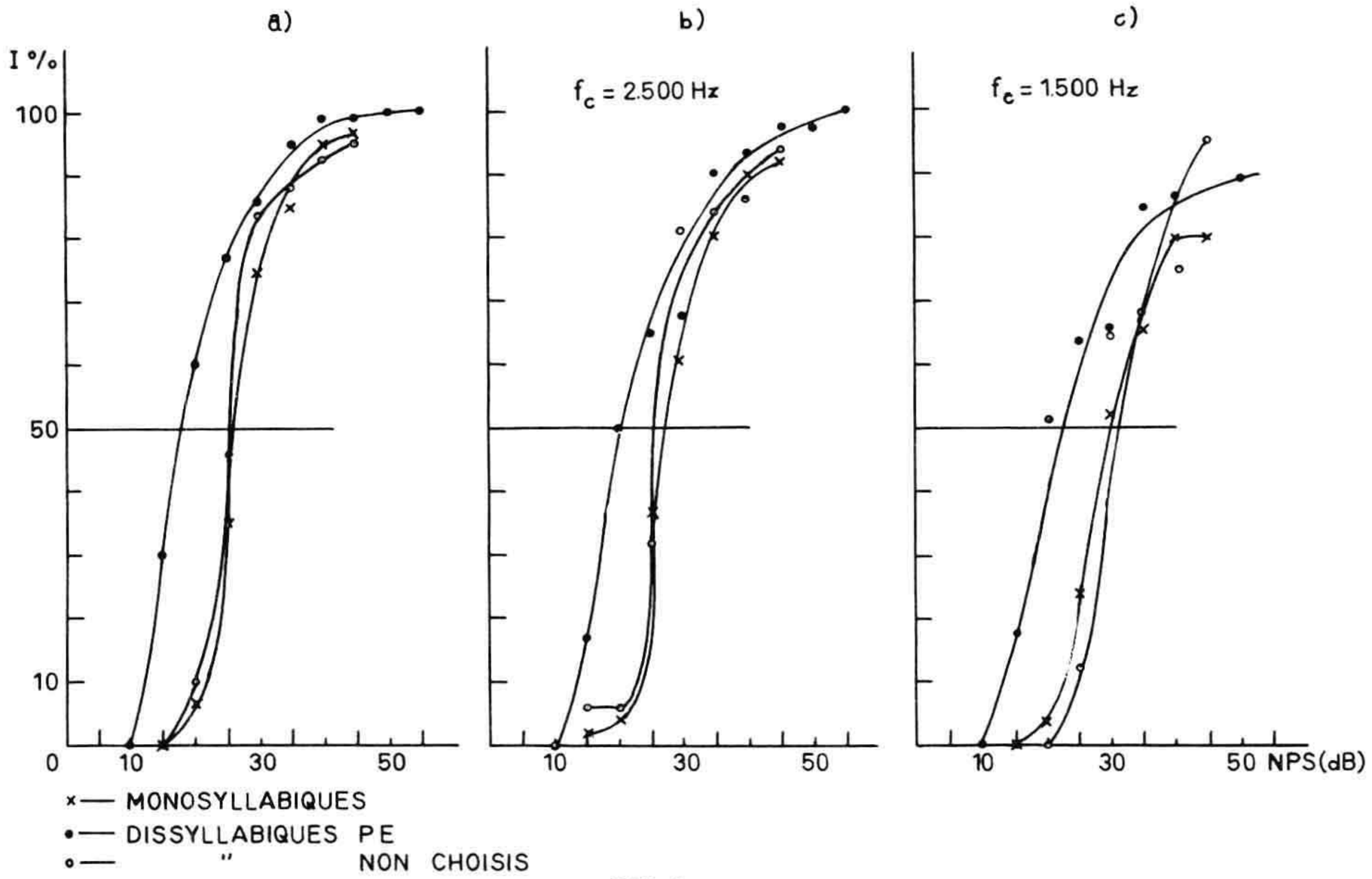


FIG. 2

hasard, puisque le contenu phonétique des deux listes présente en un certain sens des analogies, dues principalement à ce que, par le nombre de mots, réduit, une grande partie des phonèmes caractéristiques de la langue sont absents, et a la simplicité du systeme des voyelles espagnoles (5 sons de voyelle seulement).

Dans la courbe correspondant aux dissyllabiques PE, où sont représentés tous les phonèmes de l´espagnol, on obtient une grande homogénéité dans le test et la valeur du seuil d´intelligibilité, de même - que le maximun s'améliorent.

Les fig. 2b et 2c correspondent aux résultats obtenus en utilisant des filtres passebas de 2500 Hz et 1500 Hz de fréquence de coupure; on remarque l´influence de la largeur de bande dans le seuil et dans le maximun d´intelligibilité.

Dans la fig.3 on présente les résultats pour le cas d´un filtre passe bas de f_c= 500 Hz. Dans ce cas, l´information obtenue résulte pratiquement nulle et une grande distorsion se présente à partir d´un niveau de pression sonore de 60 dB.

Dans la fig.4 on donne les résultats obtenus pour essais d´intelligibilité en fonction du masque par bruit de fond. On a choisi deux sortes de sons masquants: une bande de bruit blanc de 1/3 d´octave de largeur et de 1000 Hz de fréquence centrale, et un bruit industriel de -

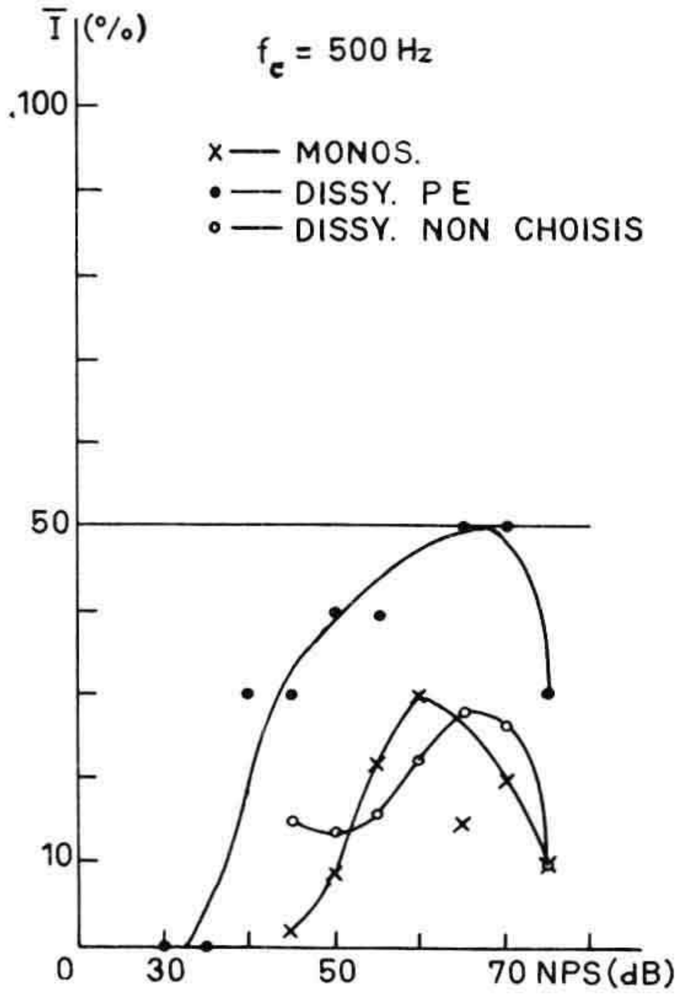


FIG. 3

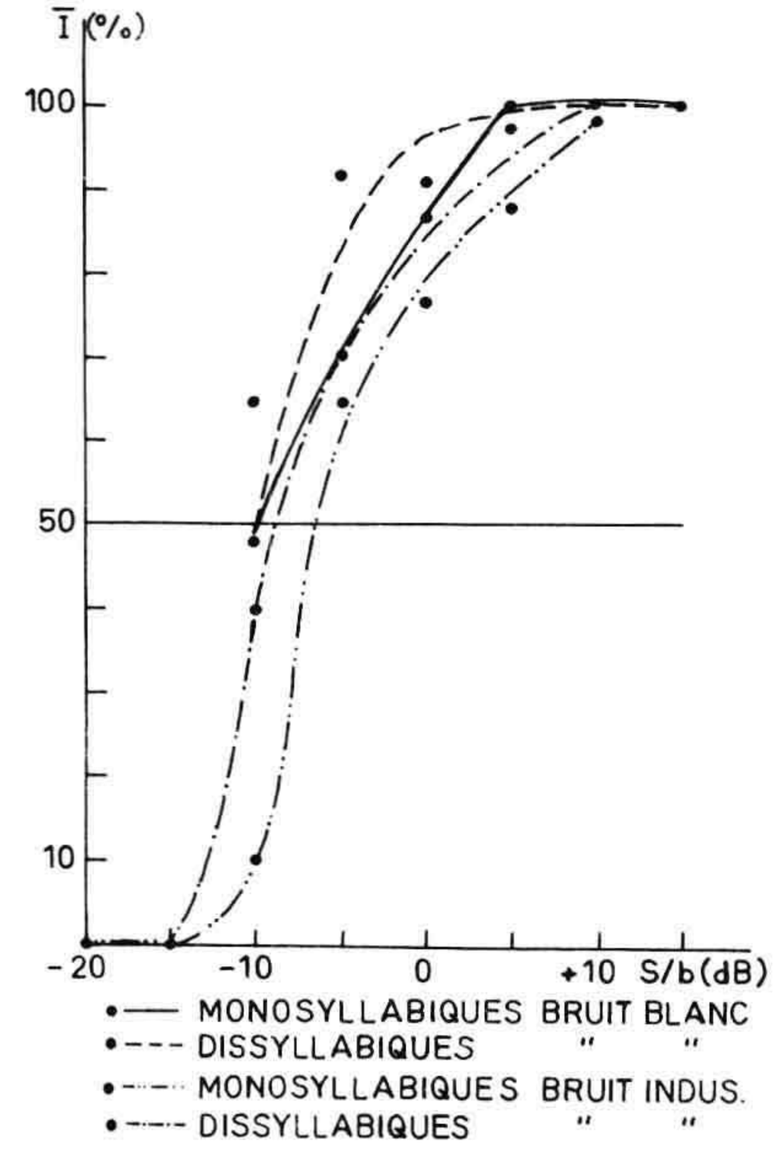


FIG. 4

spectre large(bruit de métiers à tisser).

L´effet du masque de la bande du bruit blanc, commence à être négligeable($\bar{I}$ > 97%) pour un rapport du signal au bruit, supérieur à 5 dB.

Pour le bruit industriel on parvient à des relations supérieures à 10 dB, l´effect masquant étant dans tous les cas,plus grand dans les mots monosyllabiques.

Pour une égalité du rapport du signal au bruit les taux d´intelligibilité obtenus sont toujours plus petits dans les cas où l´on emploie le bruit industriel plutôt que le bruit blanc.

3.- Conclusions

On présente un test pour épreuves d´intelligibilité en espagnol, de mots monosyllabiques et dissyllabiques, et on obtient des résultats d´intelligibilité de systèmes de transmission en fonction de la largeur de la bande, du rapport du signal au bruit et du spectre de bruit.

Bibliographie

Alarcos, F.- Fonología española; ed.Gredos, Madrid (1965)

Fletcher, H.- Speech and Hearing in Communication; D.van Nostrand Co., New York (1953)

Hirsh, I.J.- La mesure de l´audition; P. Universitaires de France, Paris (1956)

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

Influence de la distorsion harmonique d'un systeme de transmission sur l'intelligibilité.

Pfretzschner, Jaime et Jover, Adelaida
Centro de Investigaciones Físicas "L. Torres Quevedo"
Département de Acoustique
Madrid - Espagne

Introduction

Le but de ce travail est d'examiner l'influence de la distorsion harmonique sur l'intelligibilité de consonnes labiales dans la langue espagnole, d'applications dans les systemes de transmission electroacoustique. Comme technique de mesure on utilise des essais subjectifs dans le champ libre sur un auditoire typique.

L'election correcte d'un test est le point de depart. Dans notre cas on a pris 15 logatomes monosyllabiques resultants de toutes les combinations possibles des trois consonnes labiales existant dans la langue espagnole, P,B,M, avec le 5 voyelles. Afin que la voyelle qui forme une partie du logatome ne soit coupée, on a ajouté la lettre R a la fin de chacune d'entre elle. L'addition de la lettre R n'influe pas sur les resultats car son énergie est minime en face de l'energie que apporte les formants de la voyelle et n'ajoute pas d'information comme on peut l'apprecier dans le spectrogramme de la figure 1.

Sur le tableau I figure les logatomes utilisés:

TABLEAU I

MAR	MER	MIR	MOR	MUR
PAR	PER	PIR	POR	PUR
BAR	BER	BIR	BOR	BUR

Realisation.-

Une fois elues les 15 mots, on forme le test en les melangeant de forme aléatoire et on a terminé leur enregistrement a un niveau moyen de pression sonore constant.

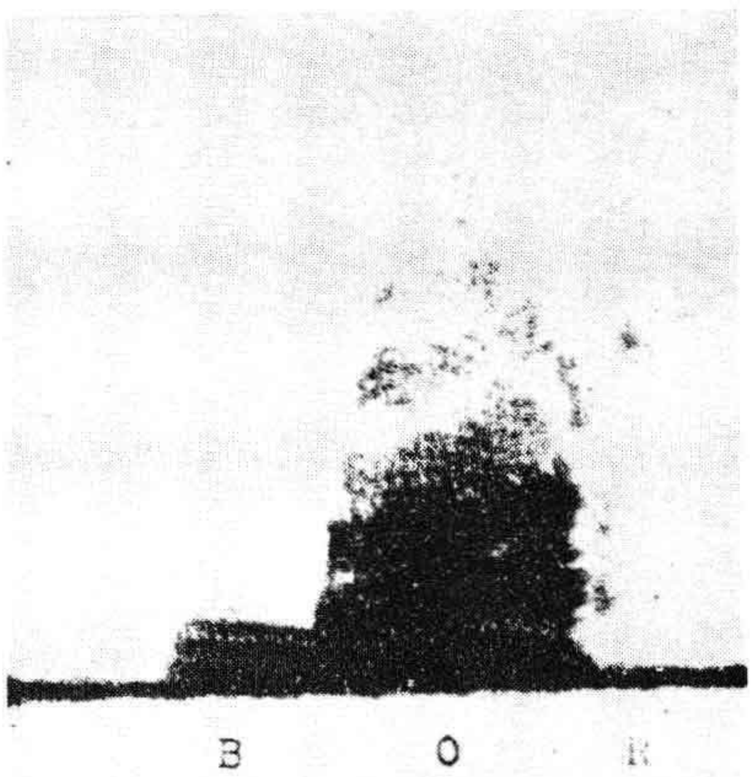


Fig.1

Comme l'objet recherché est l'examen de l'influence de la distorsion sur chacun des mots, on emploit pour obtenir cer élément distorsionateur carré qui produirait une distorsion variable a volonté et une fois elue resterait constante sur la largeur de bande requérie. Dans notre cas, l'element distorsionateur produit des pourcentage constantes de distorsion sur un intervalle de fréquences de 100 a 6000 Hz. sufisant à la marge éxigée pour une transmision de la parole.

Pour obtenir les conditions de champ libre on a utilisé la chambre anecoique du CIF L. Torres Quevedo de 250 m3. de volume et de fréquence de coupure 90 Hz. Dans le diagramme du bloc de la figure 2 on peut apprécier le systeme electroacoustique utilisé.

Dû à la limitation idiomatique du test, etant donné qui existe plus de combinations possibles et pour éviter que l'auditeur puisse se rememorer l'odre et prédire le mot qu'il allait écouter, on varie aleatoirement l'odre des mots compris dans le test dans chacune des mesures.

Sur les graphiques de la figure 3 on represente les diferentes

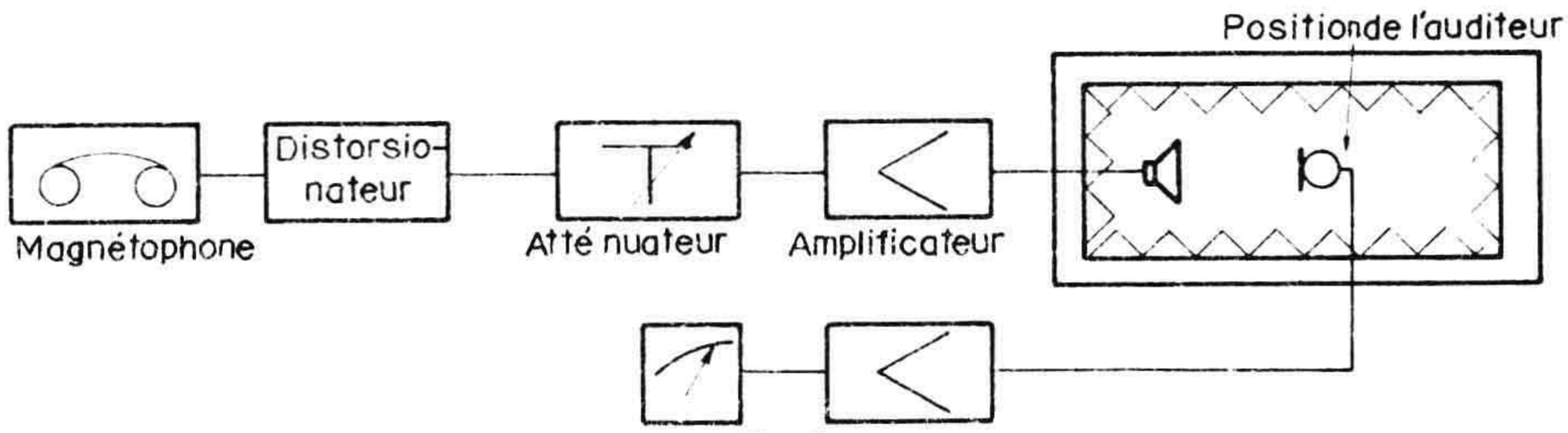


Fig.2

courbes d'intelligibilité en fonction du NPS pour des diverses pourcentages de distorsion.

Sur les figures 4 et 5 an représente le pourcentages des voyelles et des consonnes non assimilees de chacun des mots qui composent le test, en fonction du NPS et des diverse grades de distorsion.

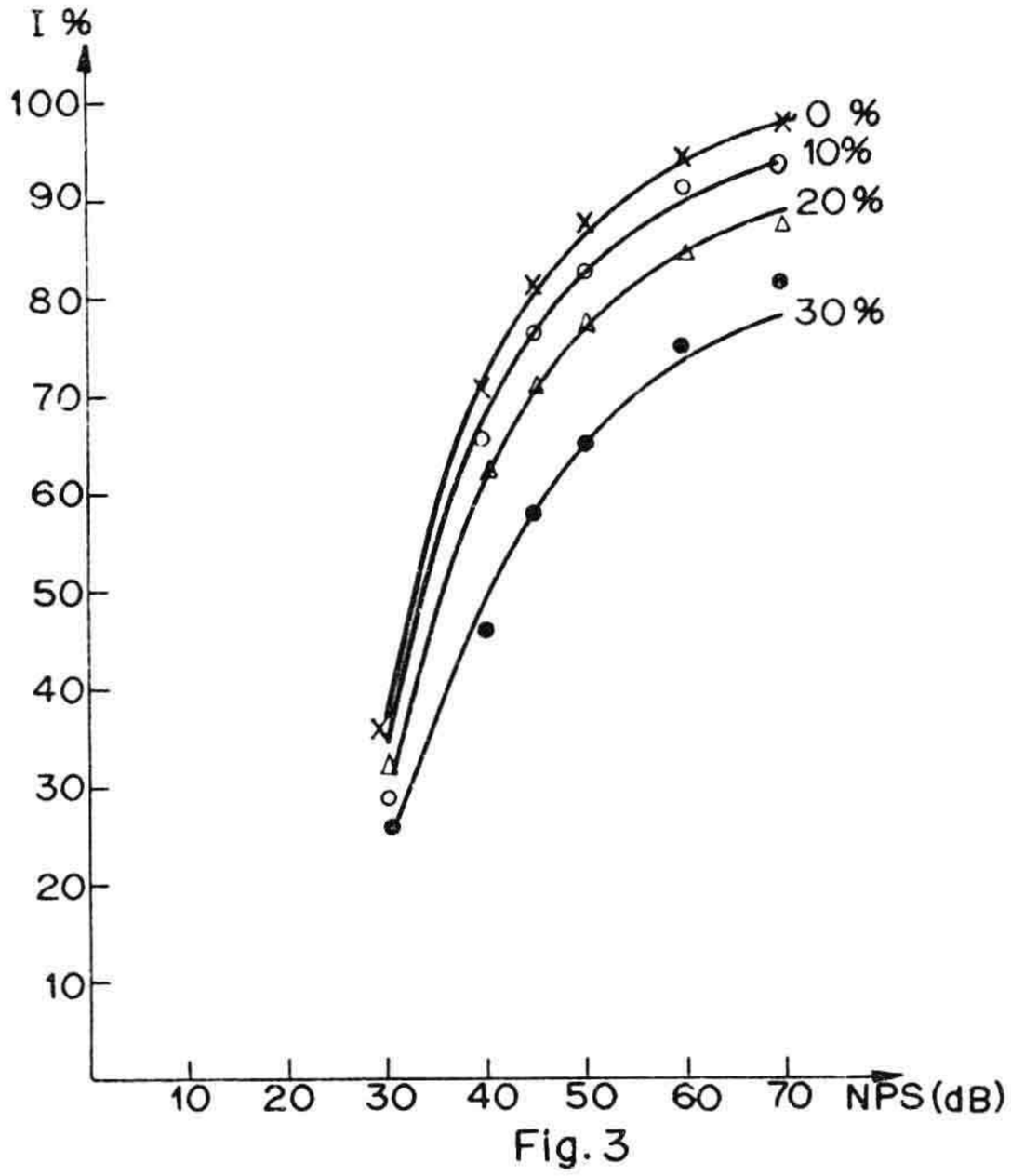


Fig. 3

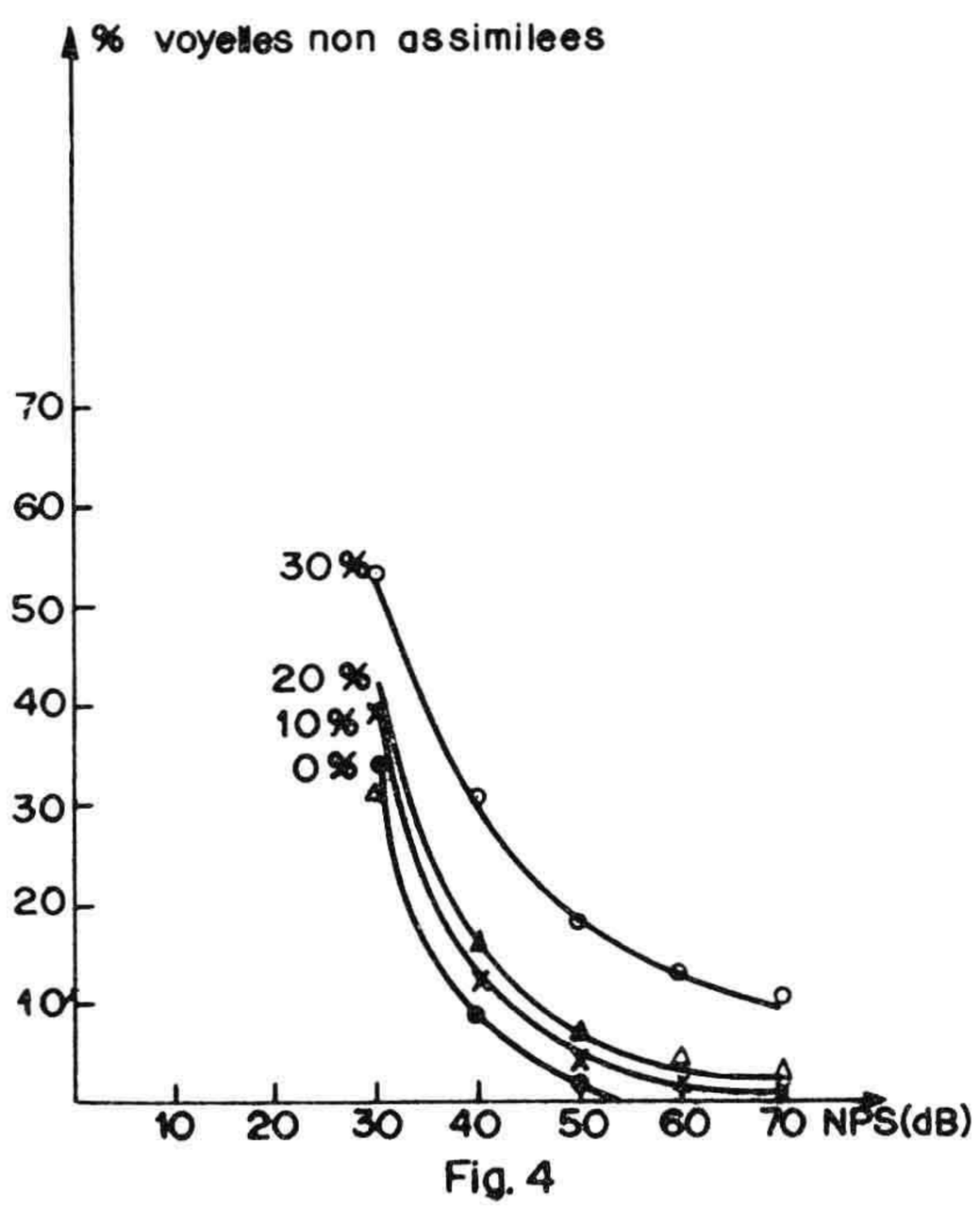


Fig. 4

Conclusion

Sur les courbes de la figure 3 on peut observer que l'influence de la distorsion sur l'intelligibilité n'est pas linéaire en fonction du niveau de pression sonore, mais parait être que pour les niveaux au-dessus des 40 dB, l'influence suit une progression geometrique de raison 2, pendant que pour des niveaux inferieurs a 40 db l'influence est lineaire, qui subit un saut en passant de 25 a 30% de distorsion.

Sur les figures 4 et 5 on observe une plus grande influence de la distorsion pour des niveaux de hautes pressions sonores qui pour de niveaux bas, autant pour des consonnes que pour de voyelles, en concordance avec l'obtenu sur la figure 3.

Dans tous les cas on appreeie que l'influence de la distorsion pour tous les niveaux, est plus grande pour le consonnes que pour le voyelles.

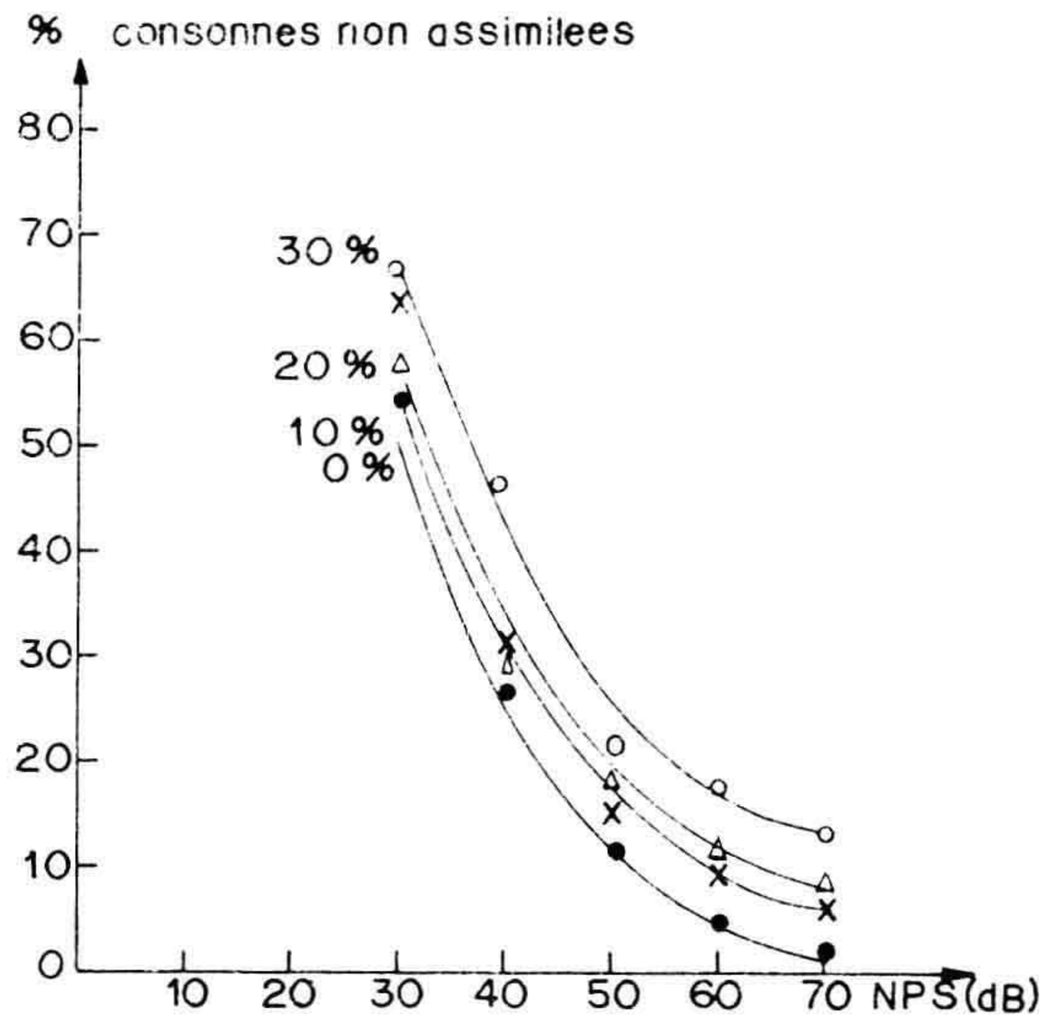


Fig. 5

De l'etude des mots non assimilees ou confondus entre eux, on observe une concordance dans leurs spectrogrammes.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

P-0-26

The Twin-Com System

J.J. Geluk

Director Research Broadcasting Laboratories NRU-NTS,
Hilversum,
Professor at the Technical University of Delft, Netherlands.

Introduction

Although in general the quality requirements for an additional sound channel in television should be kept the same as specified for the normal sound channel, in special cases one could permit some reduction in quality of both these channels. If for instance commentaries are involved, the sound quality might be regarded in the light of speech only. It is well-known that under such circumstances an audio band up to 6 kHz is adequate to ensure good intelligibility; moreover any AM Broadcast Service will not even reach this overall frequency-range while still regarded as acceptable in this respect.

This principle has led to the development of a system named the Twin-Com system. It is primarily meant to be used as a system for studio production and signal distribution, but as transmitted signal serious complications do occur in present day receivers; the most essential draw-back is common to all known systems of multiplex sound channels in television namely the "buzz" and "picture-in-sound" interferences caused by the inter-carrier sound system. On the other hand this system requires only a normal 15 kHz channel enabling to use one normal 15 kHz magnetic recording track and a normal sound distribution network. If also transmitted as such, the sound transmitter too is apt to radiate the Twin-Com signal, but even if a better system for inter-carrier reception could be found, it would remain advantageous to proceed in the studio-production and the distribution network with the Twin-Com signal, and to recode the two commentary channels to the eventually better system. Of course it could also be practicable to select one of the two commentaries out of the Twin-Com signal and to radiate the wanted sound signal in the usual way; in two-lingual countries having parts of one language only the Twin-Com distribution system presents attractive possibilities.

System description

The system, using a 15 kHz audio-channel, divides the bandwidth in a lower part up to 6,3 kHz and a single sideband signal extending from the picture line-frequency of 15.625 kHz downwards to 9 kHz. The second sound signal, also limited to 6,3 kHz is modulated as a single sideband with suppressed sub-carrier onto the video line frequency, which is not only very stable but is also available at any point in the transmission chain; in particular at the receiver or monitor this availability makes precise synchronous detection possible.

As in many existing audio transmission circuits a sharp cut-off is introduced at (video)-line-frequency, the single sideband modulator is followed by such a filter, acting as an additional single sideband filter.

The maximum level of the audio circuit has to be divided between the two sound channels, but here a remarkable property of a single sideband signal presents itself as advantageous. Limiting such a signal does not introduce harmful distorsion but merely leads to a compression of the dynamic range, so the primary sound signal level may be maintained if such compression of the second sound signal is tolerated; sometimes this effect is even appreciated for commentaries. In any case a reduction of 3 dB seems to be tolerable not so much for the signal/noise ratio, but mainly because of level fluctuations while going over from normal high-fidelity sound to the Twin-Com fashion. In fact the signal/noise ratio in the primary channel is not so much governed by the signal level and the noise, but by the level-difference of primary and secondary channel and the frequency-dividing filters used.

The level of the secondary signal has been chosen as - 7 dB relative to the level of the primary signal; thus recording of the composite Twin-Com signal is permissible with full level giving important operational advantages.

System diagram

At the production centre the electronic circuit consists out of two pré-amplifiers with associated filters, one modulator with single sideband filter and a mixing unit. The sub-carrier is available in the form of the horizontal pulses of the central picture synchronization generator. (See fig. 1a). The spectrum obtained is directly suitable for recording and distribution; the recording should always be associated with the recording of the synchronous picture in order to regenerate the sub-carrier for later demodulation. It is perhaps worth mentioning, that no special phase relation has to be maintained for detection of the second audio signal, so the circuitry remains simple.

If also "international" sound or "married" sound is involved, this should be mixed into both sound channels in order to supply a complete programme signal in both channels; if no language problem

is involved like in music programmes, the mixing should be achieved via a normal channel of the studio sound-control equipment. At the receiving point this programme then is available with full quality up to 15 kHz.

Fig. 1b gives the block diagram at the receiving point or any other point where one wants to monitor a composite Twin-Com signal. A high pass filter selects the frequency range from 9 kHz upwards, the outgoing signal of which is demodulated with a square wave synchronised with the line deflection of the picture tube. A low pass filter not only smoothes the demodulated signal but adds to the separation of the two channels of the Twin-Com signal. The primary signal is obtained after a low pass filter (cut-off frequency 6,3 kHz).

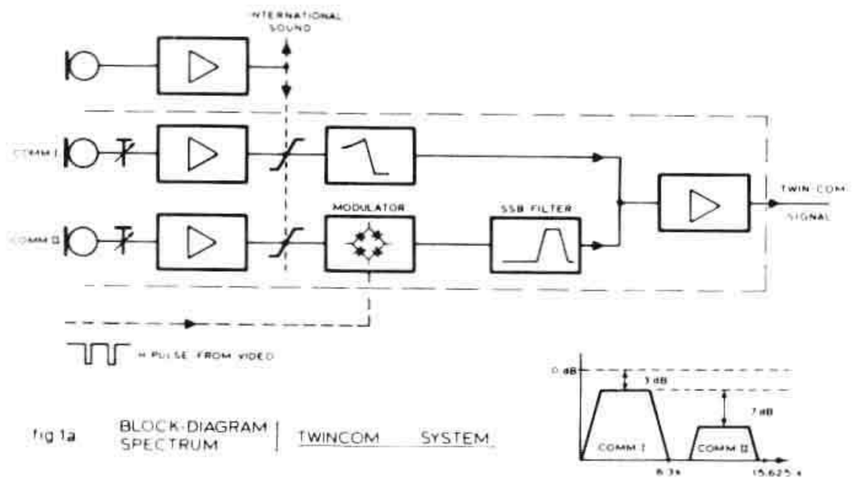


fig 1a BLOCK-DIAGRAM SPECTRUM | TWINCOM SYSTEM

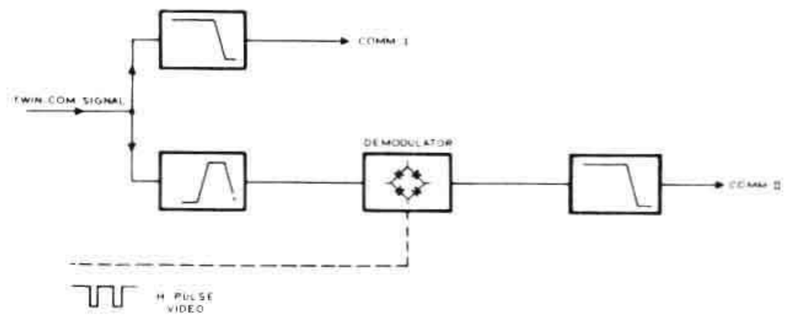


fig 1b BLOCK-DIAGRAM AT RECEIVING POINT

Some typical results

The following typical results have been obtained with laboratory equipment both in closed circuit and via an experimental transmitter. The noise whether continuous or originating from the system has been measured weighted according to the C.C.I.T.T. psophometer evaluation; the noise levels so measured with- or without the other signal being present resp. signal/noise and signal/modulation noise) give an idea of the effective separation between the two channels.

	Signal/zero noise		Signal/Noise (inclusive cross-talk)		Distorsion(k_3)	
	channel I	II	I	II	I	II
Typical tape recorder	58	40	46	38	1 %	0,6 %
Typical receiver results						
without treble reduction	50	38	34	37	2 %	1 %
with treble reduction			42	37		

The measurements in the secondary channel of the signal/noise ratios are greatly influenced by the picture content, and the receiver used; only receivers having separate I.F. circuits for the sound channel seem to comply with good signal/noise ratios for multiplex sound signals.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-4

Predictive Coding of Speech Signals

B. S. Atal and M. R. Schroeder

Bell Telephone Laboratories, Incorporated

Murray Hill, New Jersey, USA

Introduction

The problem of efficient coding of speech signals has received considerable attention over the last several decades [1][2]. While a majority of existing coding schemes can reproduce intelligible speech, the quality of the synthesized speech is often noticably inferior to that of the original speech. This degradation in speech quality is believed to be closely connected to certain operations employed in the coding process (such as voiced-unvoiced decision, pitch extraction and measurement of spectral envelope) which are either inaccurately performed or which are based on certain abstractions of the speech production and perception processes and valid only as approximations. A different approach to speech coding, which avoids these difficulties, is described in this paper.

Principle Of Predictive Coding

For efficient use of a given channel capacity it is essential to reduce the redundancy of the transmitted signal. A well known procedure for doing this is known as "predictive coding" [3], in which the redundancy is removed by subtracting from the signal that part which can be predicted from its past. The difference signal has the property, that for any fixed sampling rate, its samples are less correlated than the samples of the original signal. In general, compared to the first-order entropy of the signal the first-order entropy of the difference signal is closer to the conditional entropy (conditional on the whole past) of the original signal. The difference signal is thus better suited than the original signal for memory-less encoding (coding on a sample-by sample basis).

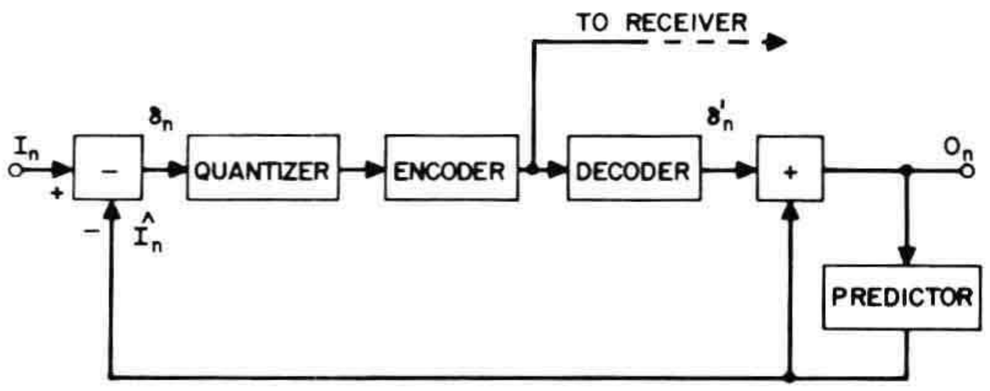


Fig. 1. Block diagram of a predictive coding system.

A block diagram illustrating the principle of predictive coding is shown in Fig. 1. The difference signal δ_n is formed by subtracting from the input signal I_n, a predicted signal $\hat{I}_n$ based on the past samples of the reconstructed signal O_n. This difference signal is quantized, encoded and transmitted to the receiver. At the same time, the transmitted signal is decoded at the transmitter and the signal reconstructed in exactly the same manner as is done at the receiver. The reconstructed signal is then used to predict the next sample of the input signal.

The theory of predictive coding systems has been extensively discussed in the

literature [3]-[5]. We should merely like to point out here one of the important features of the predictive coding. The error e_n between the reconstructed signal O_n and the input signal I_n, defined by

$$e_n \equiv O_n - I_n, \tag{1}$$

may be written as follows:

$$e_n = (O_n - \hat{I}_n) - (I_n - \hat{I}_n) = \delta_n' - \delta_n, \tag{2}$$

where δ_n is the difference signal at the quantizer input and δ_n' is the decoded difference signal appearing at the decoder output and $\delta_n' - \delta_n$ is the error introduced in the decoded difference signal by the quantizer, encoder and the decoder. Thus, according to Eq. (2), the error e_n in the nth sample of the reconstructed signal is identical to the error $\delta_n' - \delta_n$ in the nth sample of the decoded difference signal. It also follows from Eq. (2) that the "signal-to-noise" ratio of the reconstructed signal exceeds the signal-to-noise ratio of the decoded difference signal by a factor $P_I/P\delta$ where P_I and P_δ represent the powers of the input signal and the difference signal, respectively.

Efficient Predictors For Speech Signals

Two of the main causes of redundancy in speech are:

1. Quasi-periodicity during voiced segments, and
2. Lack of flatness of the short-time spectral envelope.

The redundancy due to the quasi-periodic nature of speech is removed by a predictor consisting of a delay and gain. The z-transform of the predictor is given by

$$P_1(z) = \alpha z^{-K}, \tag{3}$$

where z^{-K} represents a delay of K samples and α is a gain factor. For voiced speech, the delay K will usually correspond to a pitch period.

The redundancy caused by the spectral envelope of speech is removed by linear predictor with z-transform given by

$$P_2(z) = \sum_{m=1}^{8} a_m z^{-m}. \tag{4}$$

The reasons for using a 8th-order linear predictor are as follows: The spectral envelope of speech can be approximately represented as the product of the spectra of different "formants" and the spectral envelope of the vocal source. The optimum predictor for a single formant, limited to a finite bandwidth, is a second-order linear predictor. The spectral envelope of the vocal source is approximately proportional to ω^{-2} and can likewise be predicted by a second-order linear predictor. An 8th-order linear predictor can thus eliminate or substantially reduce, redundancies due to three formants and the spectral envelope of the vocal source. The complete predictor for speech signals combining the predictors P_1 and P_2, is shown in Fig. 2.

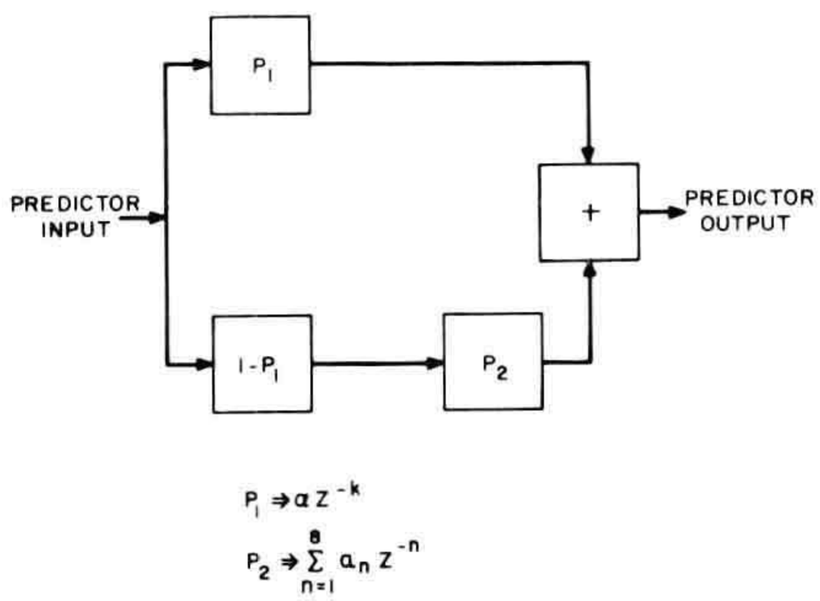


Fig. 2. Block diagram of the predictor

Determination Of Predictor Parameters

Due to the non-stationary nature of the speech signals, the predictor parameters are optimized periodically to match the characteristics of the input speech. The predictor parameters are reset every 5 msec such that the energy of the difference signal (the "non-predictable" part of the speech signal) over a given 5 msec period is minimum. The predictors P_1 and P_2 are optimized separately.

Optimization Of The Predictor P_1

The energy E_1 of the difference signal after the first prediction is given by

$$E_1 = \sum_n (S_n - \alpha S_{n-K})^2 . \tag{5}$$

The minimum value of E_1 is obtained if

$$\alpha = \sum_n S_n S_{n-K} \Big/ \sum_n S_{n-K}^2 , \tag{6}$$

and

$$\rho = \sum_n S_n S_{n-K} \Big/ \left\{ \sum_n S_n^2 \sum_m S_{m-K}^2 \right\}^{\frac{1}{2}} \tag{7}$$

is maximum. The optimum value of K was determined by direct search of the maximum of ρ. The optimum value of α was then obtained from Eq. (6).

Optimization Of The Predictor P_2

Let $\hat{v}_n$ represent the difference signal after the first prediction, with α and K set to their optimum values. The energy E_2 of the second difference signal δ_n is given by

$$E_2 = \sum_n \left[\hat{v}_n - \sum_{m=1}^{8} a_m \hat{v}_{n-m} \right]^2 . \tag{8}$$

The optimum values of the coefficients which minimize E_2 are given by

$$\underset{\sim}{a} = \Phi^{-1} \underset{\sim}{\psi} \tag{9}$$

where $\underset{\sim}{a}$ is a 8-dimensional vector whose n^{th} component is a_n, Φ is a covariance matrix with the term φ_{ij} in the ith row and j^{th} column given by

$$\varphi_{ij} = \sum_n \hat{v}_{n-i} \hat{v}_{n-j} \tag{10}$$

and $\underset{\sim}{\psi}$ is a vector with the jth component $\psi_j = \varphi_{oj}$.

Predictive Encoding System For Speech

In order to evaluate the performance of predictive coding for speech signals, the scheme shown in Fig. 3 was simulated on a digital computer. The input speech was filtered with a lowpass filter with 3-db attenuation at 3.1 kHz and an attenuation of 40 db or more for frequencies above 3.33 kHz. The sampling rate employed in the simulation was 6.67 kHz. The difference signal δ_n was quantized by a two-level quantizer with variable step size q which was optimized inside the loop to yield minimum quantization noise power. The parameters of the adaptive predictor were adjusted every 5 msec and sent to the receiver together with the quantized difference signal δ_n' and the step size q of the quantizer. The difference signal was thus transmitted at a bit rate of 6.667 kilobits per sec. No attempt was made in the study reported here to quantize the predictor parameters.

The speech signal was reconstructed at the receiver by a feedback loop containing an adaptive predictor identical to one used at the transmitter. The predictor at the receiver was also reset every 5 msec.

Results Of Subjective Tests

The speech signal produced at the receiver of the predictive coding scheme was compared by trained listeners with speech from a logarithmic PCM ("Log-PCM") encoder [6] using the same input signals, a sampling frequency of 6.667 kHz and $\mu = 100$. These tests indicate that the quality of the reconstructed speech signal is better than that of Log-PCM speech with 5 bits/sample but inferior to one with 6 bits/sample.

The reconstructed speech signal was also compared with the input speech signal contaminated by additive white noise obtained by randomly inverting the polarity of successive Nyquist samples of the input speech signal. This noise is subjectively similar to the distortion introduced by predictive coding and is therefore particularly appropriate for meaningful and reproducible comparisons. These comparisons gave an equivalent signal-to-noise ratio for the predictive coding scheme as described here of about 25 db.

References

[1] H. Dudley, J. Acoust. Soc. Am., Vol. 11, pp. 169-177, 1939.
[2] M. R. Schroeder, Proc. IEEE, Vol. 54, pp. 720-734, May 1966.
[3] P. Elias, IRE Trans. on Information Th., Vol. IT-1, pp. 16-33, March 1955.
[4] J. B. O'Neal, Jr., B.S.T.J., Vol. 45, pp. 689-721, May-June 1966.
[5] R. A. McDonald, B.S.T.J., Vol. 45, pp. 1123-1152, September 1966.
[6] B. Smith, B.S.T.J., Vol. 36, pp. 653-709, May 1957.

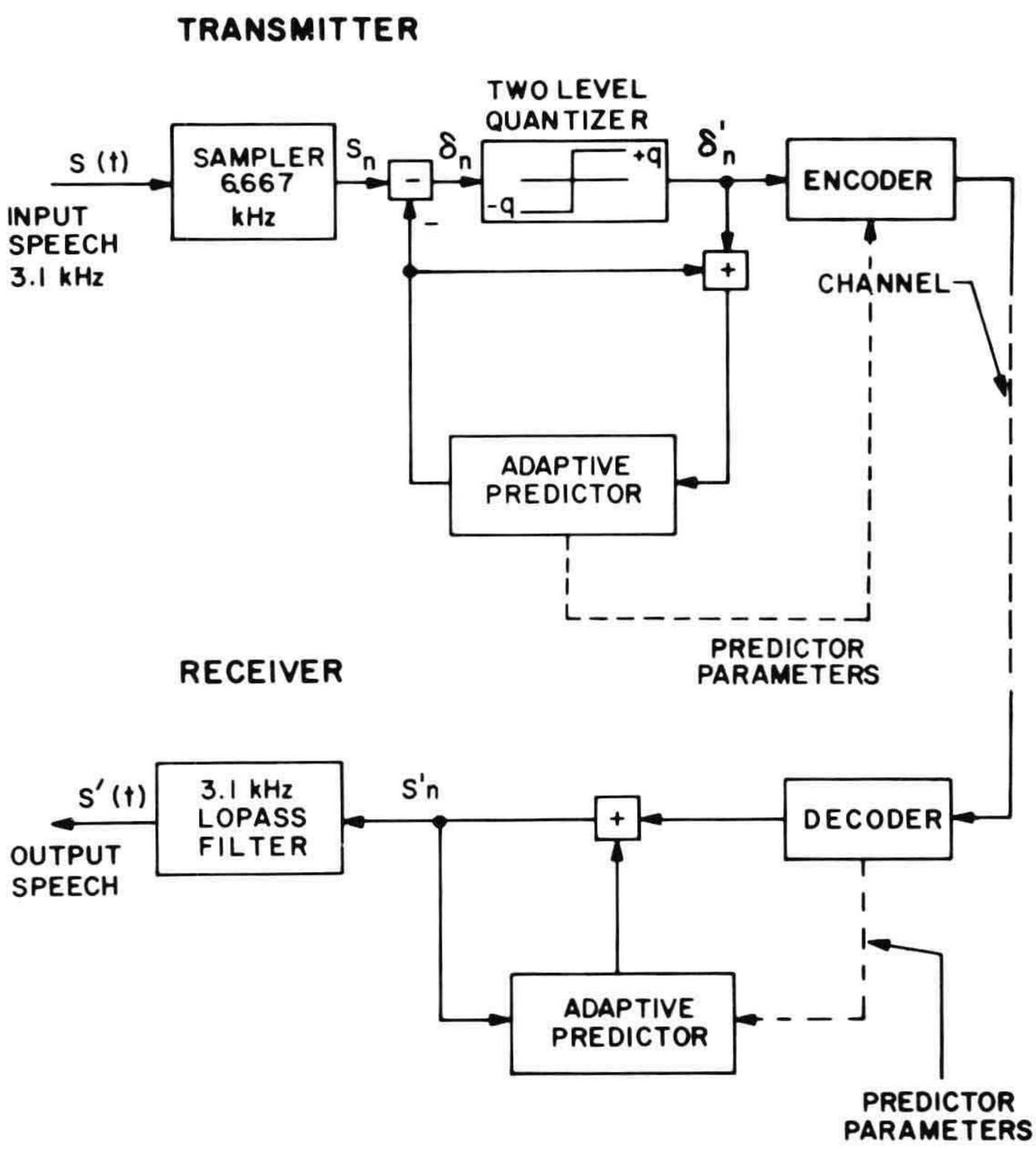


Fig. 3 Block Diagram of the Predictive Coding system for Speech

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-5

Analysis Synthesis Telephony based on the Maximum Likelihood Method

Fumitada Itakura and Shuzo Saito
Electrical Communication Laboratory, NTT
Tokyo, Japan

] Introduction

The formant vocoder is intended to achieve highly efficient transmission of speech signal, but the quality of formant vocoders is inferior to other speech compression systems due to the poor accuracy of the formant tracking.

In this paper, a more reliable method for extracting the analysis-synthesis parameters of speech signal is studied, in which the speech spectral parameters are extracted on the basis of the maximum likelihood method, and the analysis-synthesis system using the method is simulated on a digital computer.

2 Extraction of Spectral Envelope

In general, power spectral density $f(\lambda)$ of a discrete speech signal sampled at every ΔT sec could be separated into two factors; the spectral envelope and source periodicity

$$f(\lambda)=f_1(\lambda)\ f_2(\lambda) \qquad (1)$$

where $\lambda=\Delta T\omega$ is the normalized angular frequency. More specifically $f_1(\lambda)$ is a macroscopic spectral distribution resulting from the vocal tract transmission, the source wave forms, the radiation effect and etc.. On the other hand, $f_2(\lambda)$ is defined as a quotient of $f(\lambda)$ devided by $f_1(\lambda)$, so that $f_2(\lambda)$ has nearly uniform distribution, either quasi discrete or continuous according to voicing. However, because there are various fluctuations of pitch periods, source wave forms, and formant transition in a speech signal, it would be advantageous, for the extraction of spectral envelope $f_1(\lambda)$, to regard the speech signal as a random signal with continuous spectrum disregarding its quasi periodicity, that is, $f_2(\lambda)=1$. Hence we assume in this section that a short segment $X=(x_1,x_2,\ldots,x_N)$ from a sampled speech signal is a random gaussian process with the spectral density written as follows:

$$f_1(\lambda)=\frac{\sigma^2}{2\pi}\frac{1}{\left|\prod_{i=1}^{p}(1-\frac{z}{z_i})\right|^2}=\frac{\sigma^2}{2\pi}\frac{1}{\left|\sum_{i=0}^{p}\alpha_i z^i\right|^2}$$

$$=\frac{\sigma^2}{2\pi}\frac{1}{\sum_{i=-p}^{p}A_i\cos i\lambda} \qquad (2)$$

where

$$z=\exp(-j\lambda), \qquad (3)$$

$$A_i=\sum_{j=1}^{p}\alpha_j\alpha_{j+|i|} \qquad \alpha_0=1,\ \alpha_k=0 \quad (k>p) \qquad (4)$$

and the pole z_i is the i-th root of the algebraic equation with real coefficients

$$\alpha_p z^p + \alpha_{p-1} z^{p-1} + \ldots + \alpha_1 z + 1 = 0 \qquad (5)$$

Reasons for omission of the spectral zeros in Eq.(2) are [1] the human perception is not sensitive to the spectral notches, and [2] any regular spectrum can be approximated as closely as desired by Eq.(2) with increasing of p sufficiently.

Under these assumptions, a set of parameters $\tilde{\omega}=(\sigma^2, \alpha_1, \alpha_2, \ldots, \alpha_p)$, or equivalently $(\sigma^2, A_1, A_2, \ldots, A_p)$, specifies the joint probability density of X as follows

$$p(X/\tilde{\omega}) = \frac{1}{(2\pi)^{N/2}|V|^{1/2}} \exp\left(-\frac{1}{2} X V^{-1} X'\right) \qquad (6)$$

where V is the covariance matrix of X. If the sample number N in the segment is much greater than p, the logarithmic likelihood function $L(X/\tilde{\omega})$ is approximated by

$$\begin{aligned} L(X/\tilde{\omega}) &= -\frac{N}{2}\left[\log 2\pi\sigma^2 + \sigma^{-2}\sum_{i=-p}^{p} A_i \hat{v}_i\right] \\ &= -\frac{N}{2}\left[2\log 2\pi + \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\log f_1(\lambda) + \frac{\hat{f}(\lambda)}{f_1(\lambda)}\right) d\lambda\right] \end{aligned} \qquad (7)$$

where $\{\hat{v}_i\}$ and $\hat{f}(\lambda)$ are the short time correlation function and spectral density, respectively, and they are defined by

$$\hat{v}_i = \frac{1}{N}\sum_{j=1}^{N-i} x_j x_{j+i}, \quad (i = 0,1,2,\ldots,N-1) \qquad (8)$$

$$\hat{f}(\lambda) = \frac{1}{2\pi N}\left|\sum_{n=1}^{N} x_n \exp(-jn\lambda)\right|^2 = \frac{1}{2\pi}\sum_{n=-N+1}^{N-1} \hat{v}_n \cos n\lambda \qquad (9)$$

To obtain the maximum likelihood estimate of $\tilde{\omega}$, it is needed to determine $\tilde{\omega}$ which maximizes $L(X/\tilde{\omega})$ for given X or $\hat{f}(\lambda)$. This problem is mathematically equivalent to the minimization of the following functional

$$E(f_1/\hat{f}) = \int_{-\pi}^{\pi} 2[d(\lambda)+\exp(-d(\lambda))-1]d\lambda = \int_{-\pi}^{\pi}\left[d^2(\lambda) - \frac{2}{3!}d^3(\lambda) + \frac{2}{4!}d^4(\lambda) - \ldots\right]d\lambda \qquad (10)$$

where $d(\lambda)=\log f_1(\lambda)-\log \hat{f}(\lambda)$ is the logarithmic difference between the hypothetical spectrum $f_1(\lambda)$ and the short time spectrum $\hat{f}(\lambda)$. Because $E(f_1/\hat{f})$ is nonnegative and zero if and only if $f_1(\lambda)=\hat{f}(\lambda)$, it can be regarded as a measure of the goodness of fit between two spectra. In the usual error measure used in the analysis-by-synthesis spectrum matching, only the first term in Eq.(10) is considered so as to be equally sensitive to the spectral peaks and dips in $\hat{f}(\lambda)$, whereas the error measure of Eq.(10) is more sensitive to the spectral peaks and less to the dips. This is a preferable characteristic considering the tonal perception.

A set of parameters $\tilde{\omega}$ which minimizes the error measure $E(f_1/\hat{f})$ is derived from Eq.(11) and Eq.(12)

$$\sum_{i=1}^{p} \hat{v}_{|i-j|}\,\alpha_i = -\hat{v}_j \qquad (j=1,2,\ldots,p) \qquad (11)$$

$$\sigma^2 = \sum_{i=-p}^{p} A_i \hat{v}_i \qquad (12)$$

where A_i's are computed from α_i's by Eq.(4). Furthermore the digital filter with transmission function

$$T(\lambda) = \frac{1}{1+\alpha_1 z + \alpha_2 z^2 + \ldots + \alpha_p z^p} \qquad (13)$$

is always stable, and this filter is used for recovering the spectral envelope at the receiving end.

3 Extraction of Source Periodicity

To achieve higher quality reproduction, a modified correlation method for the accurate extraction of the source periodicity is proposed, in which the ill effects of the spectral envelope in the original correlation function $\{\hat{v}_n\}$ are removed.

The modified correlation function corresponding to the source periodicity is given as the Fourier coefficient of $f_2(\lambda)/2\pi$

$$w_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{f(\lambda)}{f_1(\lambda)} \cos n\lambda\, d\lambda \qquad (n=1,2,\ldots), \quad w_o = 1 \qquad (14)$$

Substituting Eq.(2),(9), and (12) for $f_1(\lambda)$, $f(\lambda)$ and σ^2, respectively, we obtain

$$w_n = \sum_{i=-p}^{p} A_i \hat{v}_{n-i} \Big/ \sum_{i=-p}^{p} A_i \hat{v}_i \qquad |w_n| \leq 1 \tag{15}$$

Since the spectral envelope of $f_2(\lambda)$ is flat by definition, the corresponding $\{w_n\}$ has sharp peaks at the delay equal to the integral multiples of the pitch period for a voiced sound, and no significant peak for an unvoiced sound. Therefore the pitch period can be easily determined by a simple peak-picking procedure. Fig.1 shows an example of the extracted spectral envelopes $f_1(\lambda)$ and the modified correlation functions $\{w_n\}$ obtained by this method. The advantages of this method are similar to the cepstrum pitch determination in many respects, except this method is carried out only in the time domain.

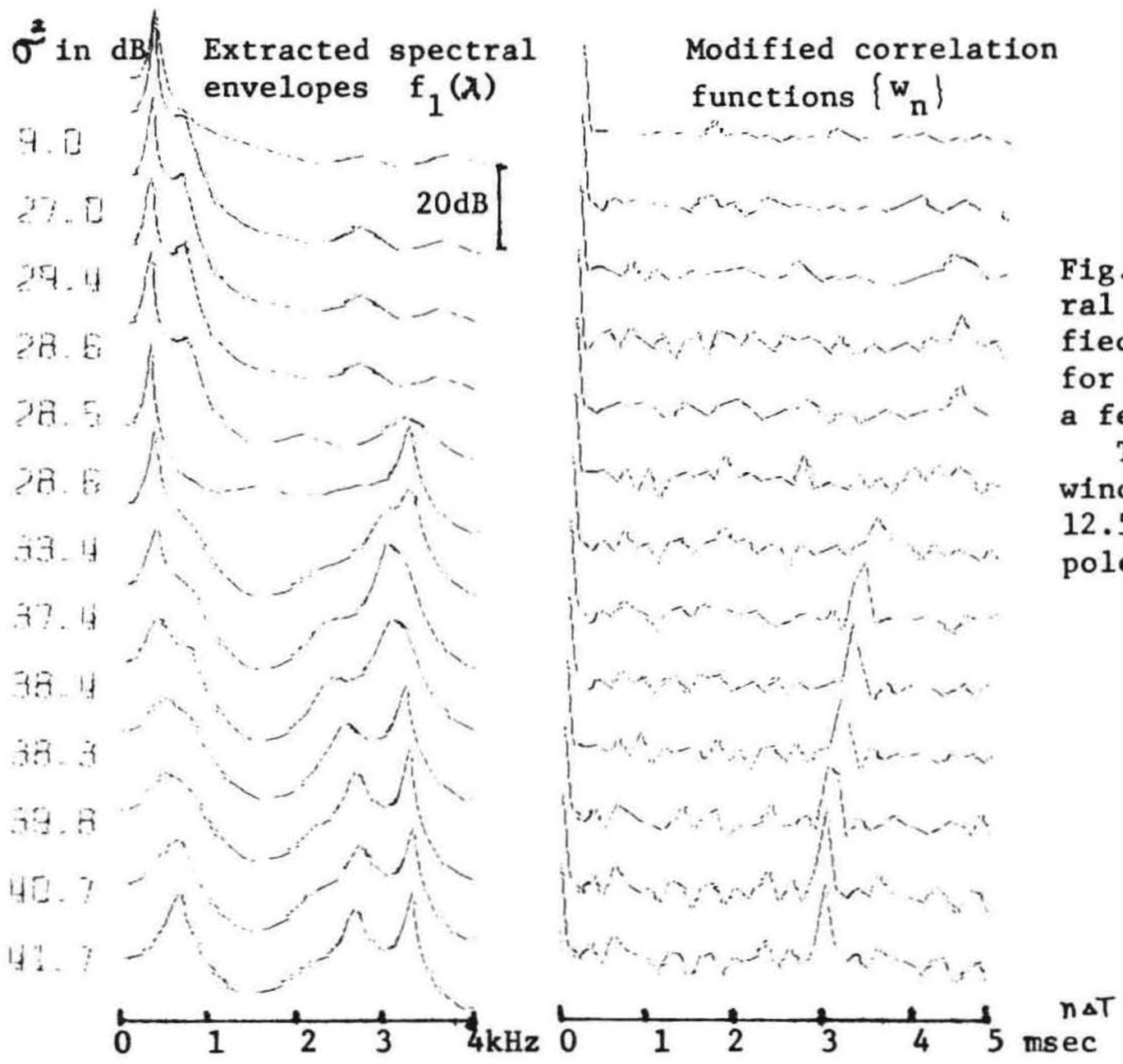


Fig.1 The extracted spectral envelopes and the modified correlation functions for a monosyllable /ge/ of a female speaker.

The 30 msec-long hamming window moved in jumps of 12.5 msec and the number of poles set to p=10.

4 Reproduction of Speech Signal

The spectral envelope of the original speech is recovered by passing the flat spectrum excitation signal through a digital filter specified by Eq.(13). The excitation signal is produced by mixing the outputs of a pulse generator and a white noise generator of equal average power in the ratio V:UV, where V is a monotonic function of the maximum w_m corresponding to the pitch period mΔT, for example,

$$V = F(w_m) = 1/1+\exp[K(\theta - w_m)] \tag{16}$$

and

$$UV = \sqrt{1 - V^2}$$

The analysis-synthesis system based on the maximum likelihood method is shown in Fig.2. Fig.3 compares the sound spectrograms of the input and output signals, when the sampling period for the control signals is set to 10 msec. The total bit rate required for the control parameters is about 5,400 bits per second for the sampling rate 50 per second, allowing 9 bits per sample for the 10 spectral envelope signals and 6 bits per sample for the 3 excitation signals. This rate is about one tenth of that required for the conventional PCM system.

5 Conclusion

The effectiveness of the analysis-synthesis system was confirmed by computer simulation. Some advantages of this system over the conventional formant vocoders are its higher analyzing accuracy in extracting the spectral envelope and the source periodicity and the possibility of completely eliminating the filter banks for the analysis and synthesis of speech spectra.

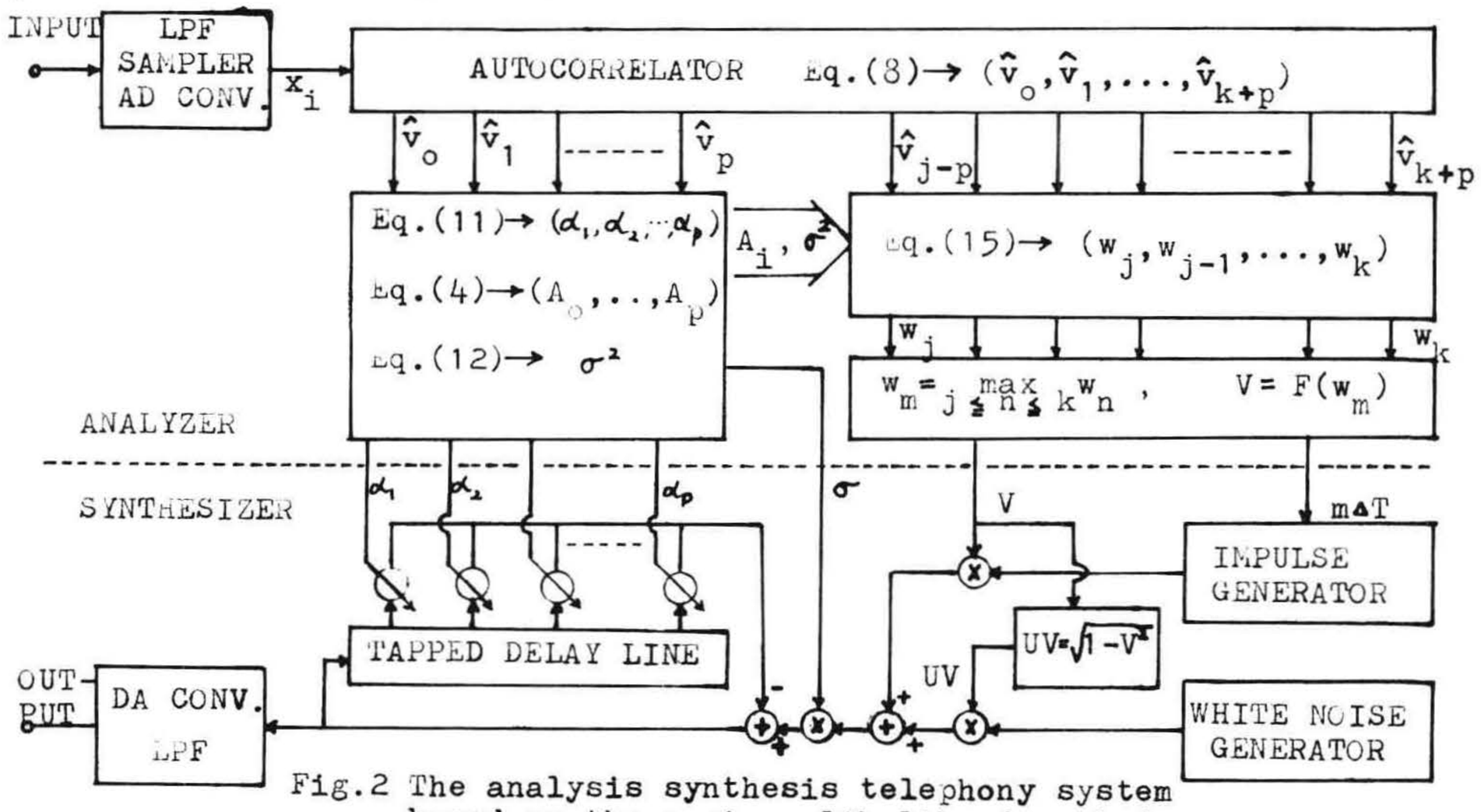


Fig.2 The analysis synthesis telephony system based on the maximum likelihood method

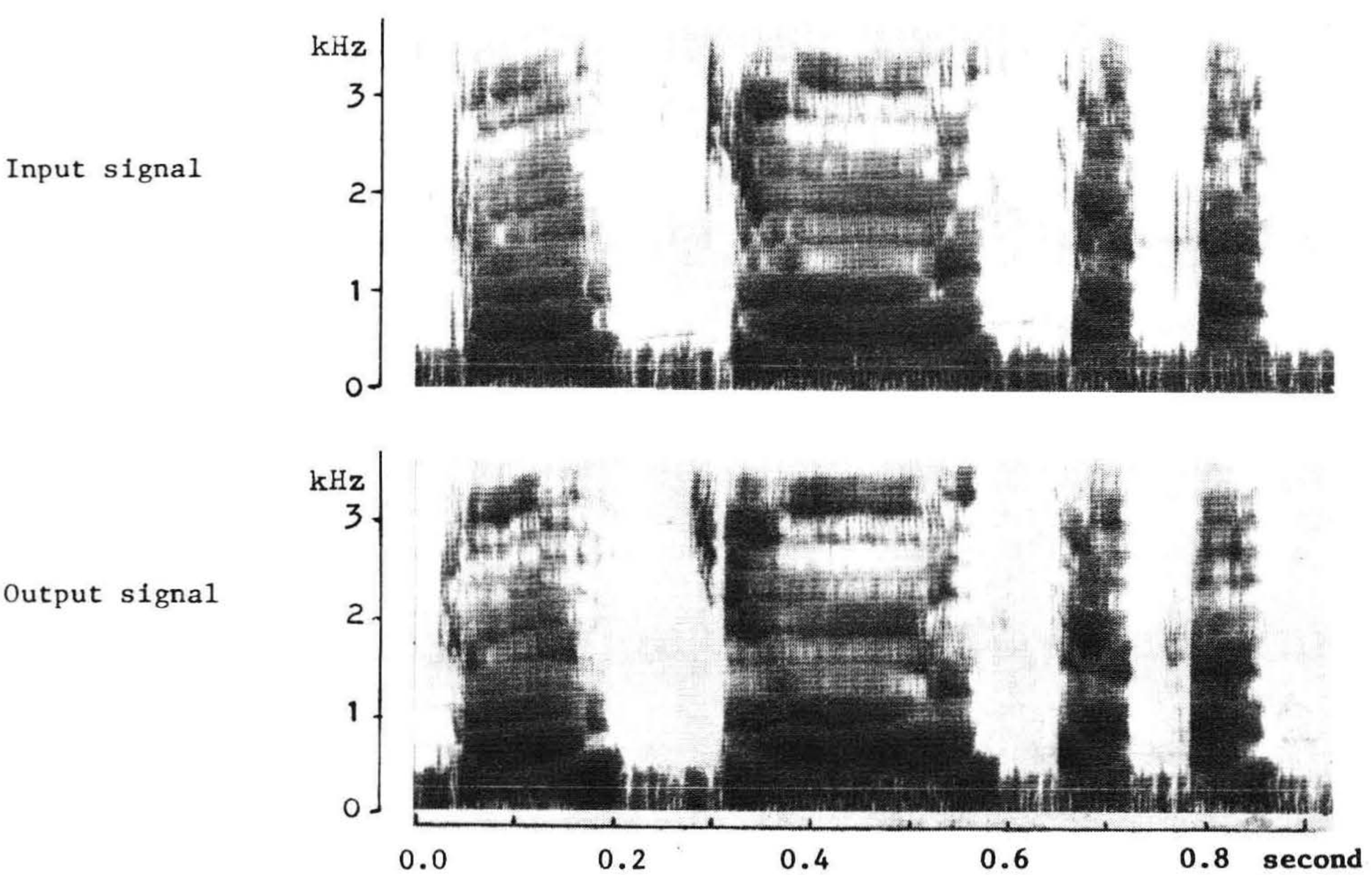


Fig.3 Sound spectrograms of input and output signals of the analysis-synthesis system. ΔT=0.]25 msec, p=]0, θ=0.2 and K=20.0

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-6

LA SYNTHESE DE LA PAROLE A PARTIR DE DIGRAMMES PHONETIQUES

LEIPP (E) Maître de recherche au CNRS
CASTELLENGO (M) ; LIENARD (J.S.)
Laboratoire d'Acoustique de la Faculté des Sciences de Paris*

1°) INTRODUCTION . Une recherche systématique sur la structure physique des signaux acoustiques en général nous a conduit à nous intéresser aux problèmes de la parole.Une approche originale du problème nous a permis d'aboutir à une synthèse de parole intelligible avec des moyens très limités.Notre procédé est à présent suffisamment au point pour que nous puissions en donner les grandes lignes.

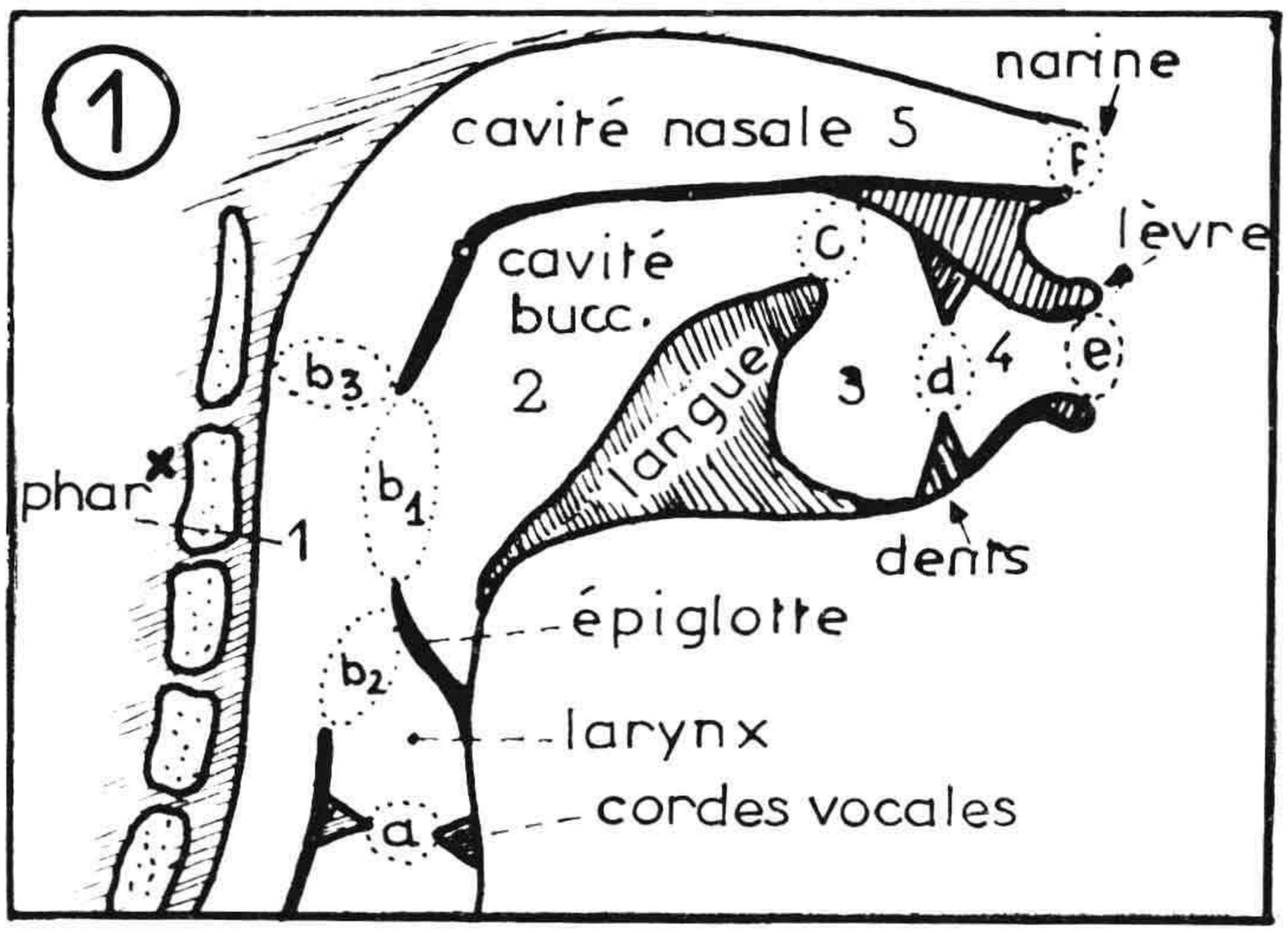
2°) POSITION DU PROBLEME . La parole normale est un ensemble de signaux acoustiques chargés arbitrairement de signification,et qui sont destinés à communiquer des messages dans le milieu naturel de l'homme.Elle s'est adaptée à ce milieu qui implique la présence d'une grande variété de bruits de fond essentiellement évolutifs et susceptibles de détruire plus ou moins les signaux.On comprend dès lors la nécessité d'une redondance du signal physique de la parole.Mais si on se place dans des conditions particulières où le bruit de fond n'intervient plus,cette redondance représente une perte inutile d'énergie.Il était donc intéressant de chercher à définir cette redondance de façon précise,puis de tenter de l'éliminer pour ne conserver que le squelette informatif,c'est à dire le minimum nécessaire et suffisant du signal.Nous y avons largement réussi, et,du même coup,nous avons simplifié le problème de la synthèse de la parole au point de pouvoir la réaliser avec des moyens très économiques.

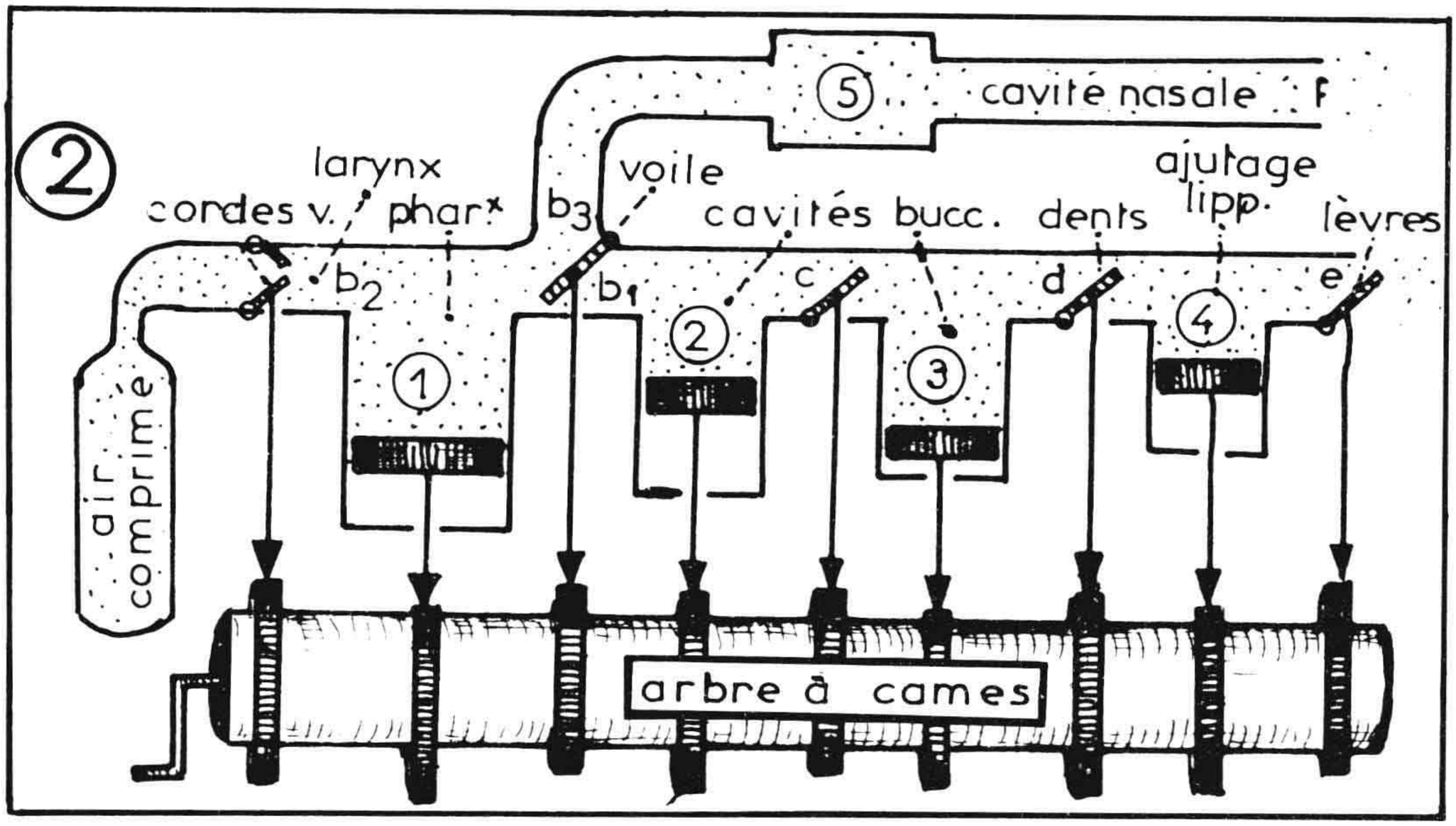
Pour réduire au maximum la redondance,nous avons commencé par n'utiliser que la bande téléphonique (entre 100 et 4000 Hz),dont l'expérience montre qu'elle est suffisante.D'autre part,nous avions observé que le spectre de raies délivré par les cordes vocales et contenant la plus grosse partie de l'énergie acoustique de la parole,n'avait d'autre utilité que de faire porter la voix au loin.La meilleure preuve en est qu'on peut le supprimer en déconnectant les cordes vocales par un artifice très simple:il suffit d'utiliser la voix chuchotée dont on vérifie que l'intelligibilité est totale.

* Département du Laboratoire de Mécanique Physique de la Faculté des Sciences de Paris (Prof. SIESTRUNCK) .

3°) MECANISME ET STRUCTURE PHYSIQUE DE LA PAROLE CHUCHOTEE. L'appareil phonatoire normal est constitué d'une série de cavités couplées,de volume et d'ouvertures réglables à volonté dans une large mesure .(Fig 1). Il s'agit d'une véritable machine à faire des sons dont on peut aisément réaliser un modèle mécanique de fonctionnement (Fig 2) .Le mouvement compliqué de l'appareil phonatoire normal lors de l'articulation d'un mot est simulé ici grâce à un arbre à cames qui synchronise les mouvements élémentaires des cavités et ouvertures.Chaque came représente un sous-programme.Si l'on excite ce dispositif en insufflant de l'air comprimé on produit des bruits de bande évolutifs entrecoupés d'explosions plus ou moins brèves,colorées par les cavités voisines de l'explosion.

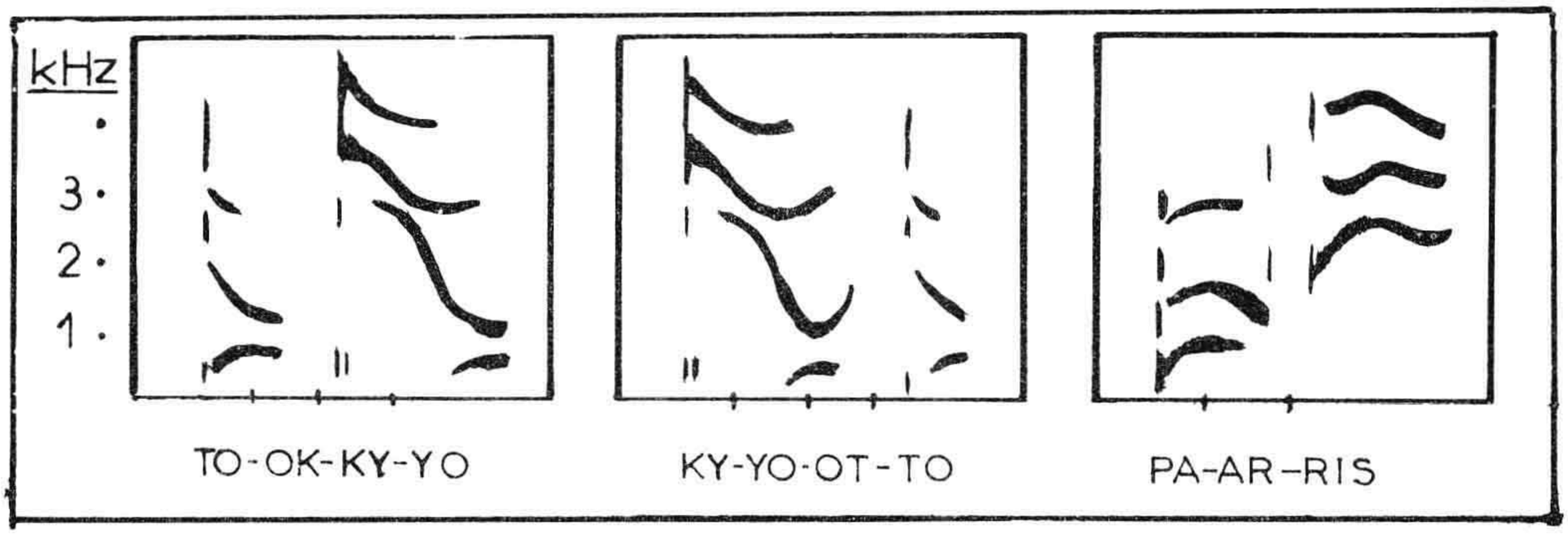
Du point de vue acoustique,le programme de l'arbre à cames se traduit donc par un signal acoustique complexe à trois dimensions:temps, fréquence,niveau. L'expérience montre que l'une de ces grandeurs, le niveau,est d'importance secondaire du point de vue du contenu sémantique du signal. C'est pourquoi la représentation sonagraphique du phénomène est hautement significative.En fait,le sonagramme véhicule l'essentiel de l'information.





LA SYNTHESE DE LA PAROLE A PARTIR DE DIGRAMMES PHONETIQUES.

Le sonagramme d'un mot chuchoté se présente sous l'aspect d'un graphisme mettant en évidence une <u>forme globale</u>,originale pour chaque mot,et,par conséquent parfaitement reconnaissable dans chaque cas sans ambiguïté. Par exemple,on ne confondra jamais "TOKIO" avec "KIOTO" ou "PARIS",pas plus qu'on ne confond le dessin d'un chat et d'une maison.(fig 3)



Comprendre un mot,c'est donc capter,reconnaitre et décoder une forme acoustique;celle-ci apparait le plus clairement possible dans la parole chuchotée,mais se retrouve bien entendu dans la parole normale.Nous avons longuement étudié cette question et observé que chaque forme globale pouvait être décomposée en éléments raccordables.Chacun de ceux-ci correspond non à un phonème,mais au <u>mouvement</u> de l'appareil phonatoire entre <u>deux</u> phonèmes voisins,en bref à un sous-programme de la machine à parler (mouvement d'une,ou de plusieurs cames simultanément).Un mot n'est donc pas décomposable phonétiquement en phonèmes,mais en éléments phonétiques qui sont des digrammes (association de deux phonèmes) et que nous appelons "phonatomes".Ainsi le mot "TOKIO" n'est pas la somme de cinq phonèmes mais l'enchainement de quatre phonatomes: TO OK KI IO ,qui se raccordent par définition,donnant de la parole fluide.

En français on admet l'existence d'environ 30 phonèmes principaux; leur association deux à deux donnera donc 900 digrammes,dont nous avons réalisé le dictionnaire complet.On notera que la forme de beaucoup de digrammes est réversible dans le temps ("iou" inversé donne "oui").Il suffit en fait de quelque 400 digrammes pour dessiner la forme globale de la totalité des mots utilisés.On dispose ainsi d'une véritable "sténographie acoustique" dont l'intérêt est évident puisqu'on peut,<u>sans analyse préalable</u> d'un mot,en dessiner la forme graphique.

Pour vérifier le bien fondé de nos hypothèses de départ,nous avons construit un synthétiseur du type play-back,l'ICOPHONE,qui est exactement l'inverse du sonagraphe.Tout graphisme est lu par un groupe de 44 photodiodes et transformé en signal acoustique global instantanément audible.

Nous avons vérifié avec une centaine de sujets que les digrammes de notre dictionnaire étaient auditivement reconnaissables;les mots entiers et les textes réalisés sont intelligibles à fortiori,et,point important, pour des personnes non suggestionnées,quoiqu'il s'agisse d'une voix chuchotée assez particulière,similaire à une bonne voix de laryngectomisé. Insistons sur le fait que nous n'avons pas cherché à reproduire fidèlement la voix normale,mais à réaliser une voix intelligible aux moindres frais. Nous avons à présent une assez longue expérience en ce domaine:dictionnaire et méthode sont suffisamment au point pour que nous ayons pu envisager d'entrer nos résultats en machine.

4°) SYNTHESE DE LA PAROLE A L'ORDINATEUR (*). Le digramme phonétique, graphisme simple,est facilement quantifiable et numérisable;d'où l'idée de les utiliser pour faire parler un ordinateur.L'intérêt de l'opération est évident:manuellement,avec l'ICOPHONE,il faut quelque 10 secondes pour dessiner un digramme,soit environ 40 secondes pour écrire "TOKIO".Pour peu que l'on mette en mémoire,comme sous-programmes,les 900 digrammes du dictionnaire,il doit être possible alors de faire parler la machine pratiquement en temps réel,l'appel d'un digramme étant pratiquement instantané.Cet appel peut être commandé par clavier (genre sténotype) ou par programme séparé.

La quantification des digrammes terminée,il a fallu construire un organe de sortie "son" approprié.Les résultats auditifs sont bien entendu exactement comparables à ceux que nous avions obtenus par voie manuelle avec l'ICOPHONE;mais le gain de temps est extraordinaire et d'ores et déjà nombre d'applications sont apparues que nous envisageons de développer par la suite.

5°) CONCLUSIONS . En partant d'une idée originale,celle de l'élimination de la redondance acoustique de la parole,nous avons été conduits à étudier la structure physique de la parole chuchotée.Sur un diagramme fréquence-temps (sonagramme),celle-ci se présente sous l'aspect de graphismes simples,de formes globales que l'on peut découper en éléments dont chacun correspond à un digramme phonétique (association de deux phonèmes).La combinatoire entre les 900 éléments phonétiques du français,permet de reconstituer sans analyse préalable n'importe quel mot ou discours en parole intelligible.Pour d'autres langues (allemand,anglais etc),l'adjonction d'un très petit nombre de digrammes suffit pour arriver au même résultat.La lecture des formes représentant les mots est faite par un playback que nous avons construit ,l'ICOPHONE.Le dictionnaire complet des digrammes phonétiques a été quantifié et numérisé pour entrée en ordinateur;ce dernier est ainsi capable de parler pratiquement en temps réel, grâce à une sortie son spécialement construite à cet effet.De nombreuses applications sont à l'étude.

BIBLIOGRAPHIE

LEIPP (E); CASTELLENGO (M) ; LIENARD (J.S) . PAROLE ET GESTALTTHEORIE
Colloque sur la parole (GALF) . Lannion Juin 1965
Tirage interne de la Faculté des Sciences

LEIPP (E) INFORMATION SEMANTIQUE ET PAROLE
Bulletin G.A.M N°22 Paris Juin 1966
Edition interne Faculté des Sciences de Paris.

LEIPP (E) ; CASTELLENGO (M) ; LIENARD (J.S) ; SAPALY (J)
Structure physique et contenu sémantique de la parole
Colloque du GALF.Grenoble (Avril 1967)
A paraitre dans la Revue Française d'Acoustique

LEIPP (E) LE CONTENU INFORMATIF DE LA PAROLE
Comptes rendus du IV° Congrès d'Acoustique de Budapest(1967)

LEIPP (E) MECANIQUE ET ACOUSTIQUE DE L'APPAREIL PHONATOIRE
Bulletin G.A.M. N°32 Paris d cembre 1967
Edition interne Faculté des Sciences de Paris.

(*) Etude conduite en collaboration avec le Centre de Calcul Analogique du Centre National de la Recherche Scientifique,dans le cadre des recherches sur le calcul hybride (Prof. L. MALAVARD) .

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-7

ZEX Coding for Speech Transmission

Shuzo Saito and Kiyoshi Hashimoto*

Electrical Communication Laboratory, NTT
Tokyo, Japan

Introduction

There are several speech coding transmission systems extracting the information bearing element of speech wave. Among them, the extremal coding system[1], proposed by M. V. Mathews, is a much interesting one and its resulting speech has fairly good quality. But it seems that there is the inherent upper bound of tonal quality in this system. A signal wave as shown in Fig. 1 will be reconstructed as a dashed line in the extremal coding. It is seen that there are several shifts at the zero crossing points of speech wave.

It is well known that much speech information is preserved at the zero crossing point of speech wave[2]. Then the deviation at the zero crossing point of speech wave may result in the degradation of reconstructed speech. It seems that there will be one possibility to overcome the upper bound of tonal qualty in the extremal coding by superposition of the zero crossing information to the extremal coding. Such a superposed speech coding system is named as the ZEX coding system.

Digital Simulation of ZEX Coding

In the ZEX coding system, informations transmitted are the times of zero crossing and extreme, and the amplitude of extreme of speech wave.

Extractions of both zero crossing and extreme points are easily achieved by the well known methods, that is, ones related to the infinite peak clipping of original speech wave and its derivative respectively.

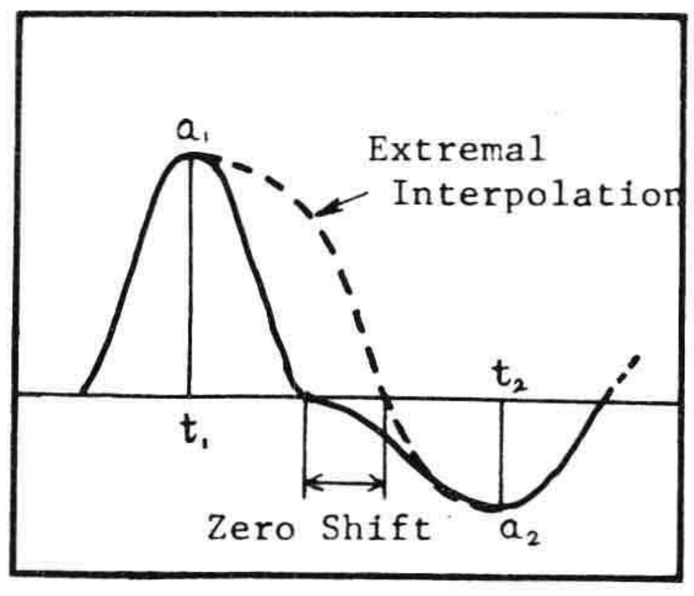


Fig. 1 Shift in the zero crossing point

* Present address is the University of Electro-Communications, Chofu, Tokyo, Japan

Using these methods, we get a temporal sequence of zero crossing and extreme points as shown in Fig. 2. It is seen that there are three kinds of sequential pairs, that is, (1) zero--extreme, (2) extreme--zero and (3) extreme--extreme.

To reconstruct speech wave from the transmitted information of zero crossing and extreme points, following interpolations are made,

(1) for zero $(0, t_i)$--extreme (a_{i+1}, t_{i+1}) pair

$$S'(t) = a_{i+1}\left\{1-\left(\frac{t_{i+1}-t}{t_{i+1}-t_i}\right)^2\right\} \quad (1)$$

(2) for extreme (a_i, t_i)--zero $(0, t_{i+1})$ pair

$$S'(t) = a_i\left\{1-\left(\frac{t-t_i}{t_{i+1}-t_i}\right)^2\right\} \quad (2)$$

(3) for extreme (a_i, t_i)--extreme (a_{i+1}, t_{i+1}) pair

$$S'(t) = a_i+(a_{i+1}-a_i)F\left(\frac{t-t_i}{t_{i+1}-t_i}\right) \quad (3)$$

where $F(X)=X^2(3-2X)$ (4)

in which $S'(t)$ is the reconstructed speech and $t_i<t<t_{i+1}$.

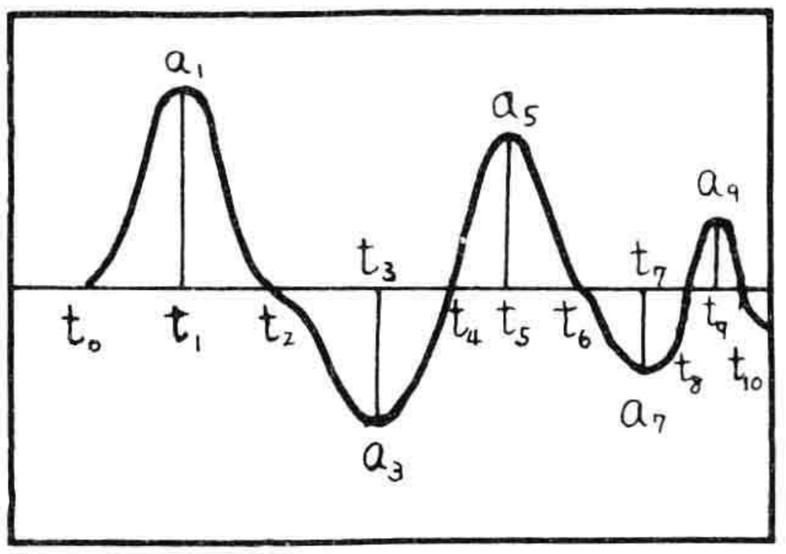


Fig. 2 Temporal sequence of zero crossing and extreme points

Block diagram of the digital ZEX coding system is shown in Fig. 3. Low pass filtered speech wave is fed to the sampling circuit and also to the type selector. In the first stage of the latter, the times of zero crossing and extreme points are detected from speech wave. Then the control signals are sent to the timing coder and the amplitude one and finally, the type of the sequential pair of zero and extreme is determined, coded and sent to the code transmitter.

Time interval between the sequential pair is calculated and quantized to 8 times of sampling frequency in the timing coder and sent to the code transmitter. Amplitude at the extreme point is calculated by a interpolation method and quantized to 7 bits and then sent to the code trans-

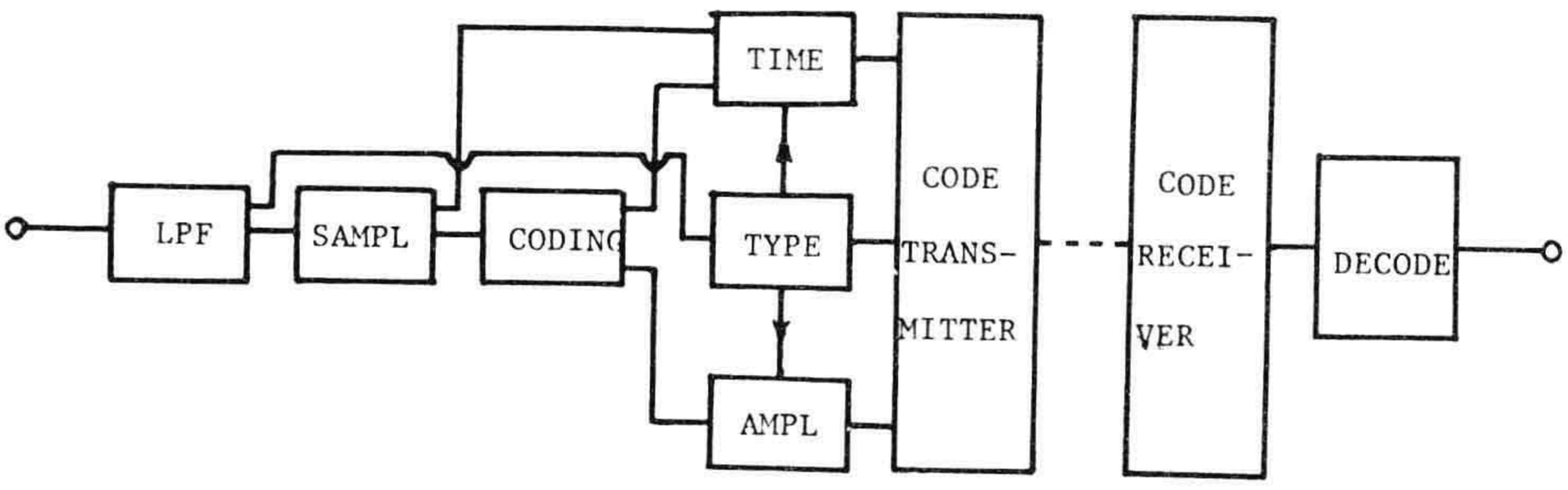


Fig. 3 Block diagram of the ZEX coding system

mitter are arranged in series and sent to the receiving end. In receiving end speech wave is reconstructed by use of the equations (1), (2) and (3).

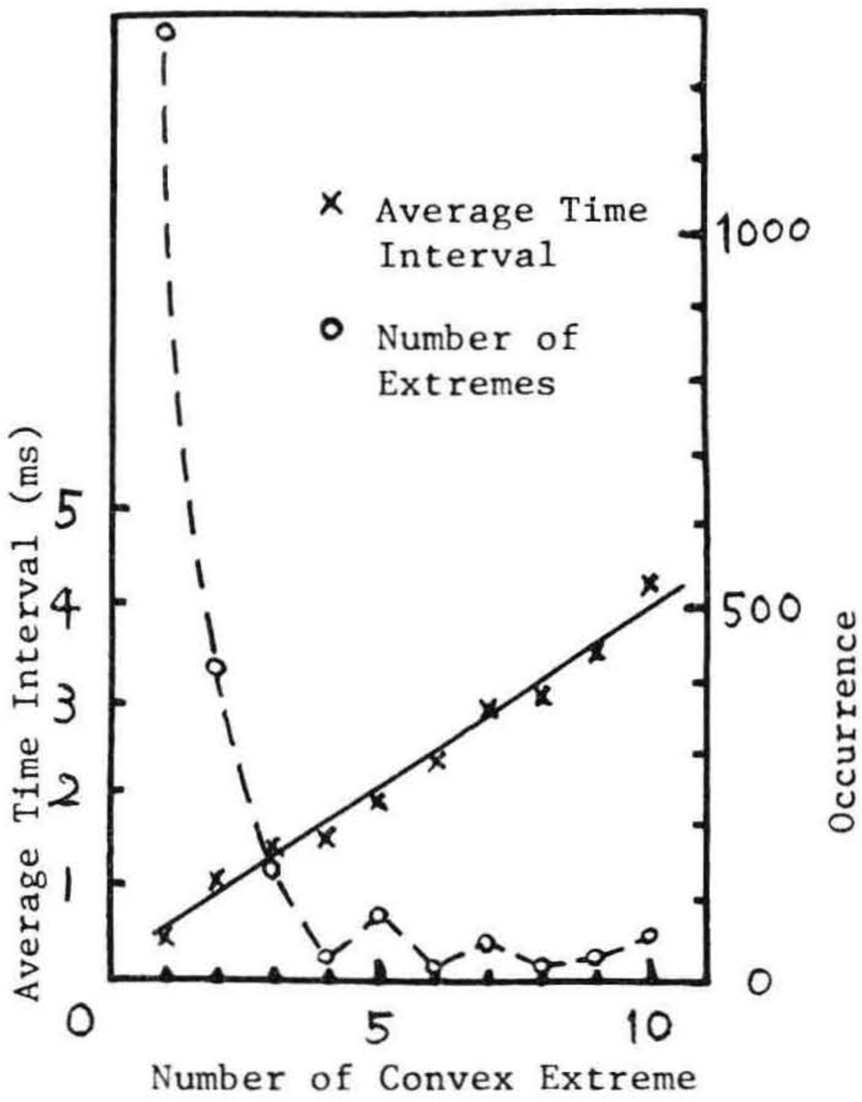
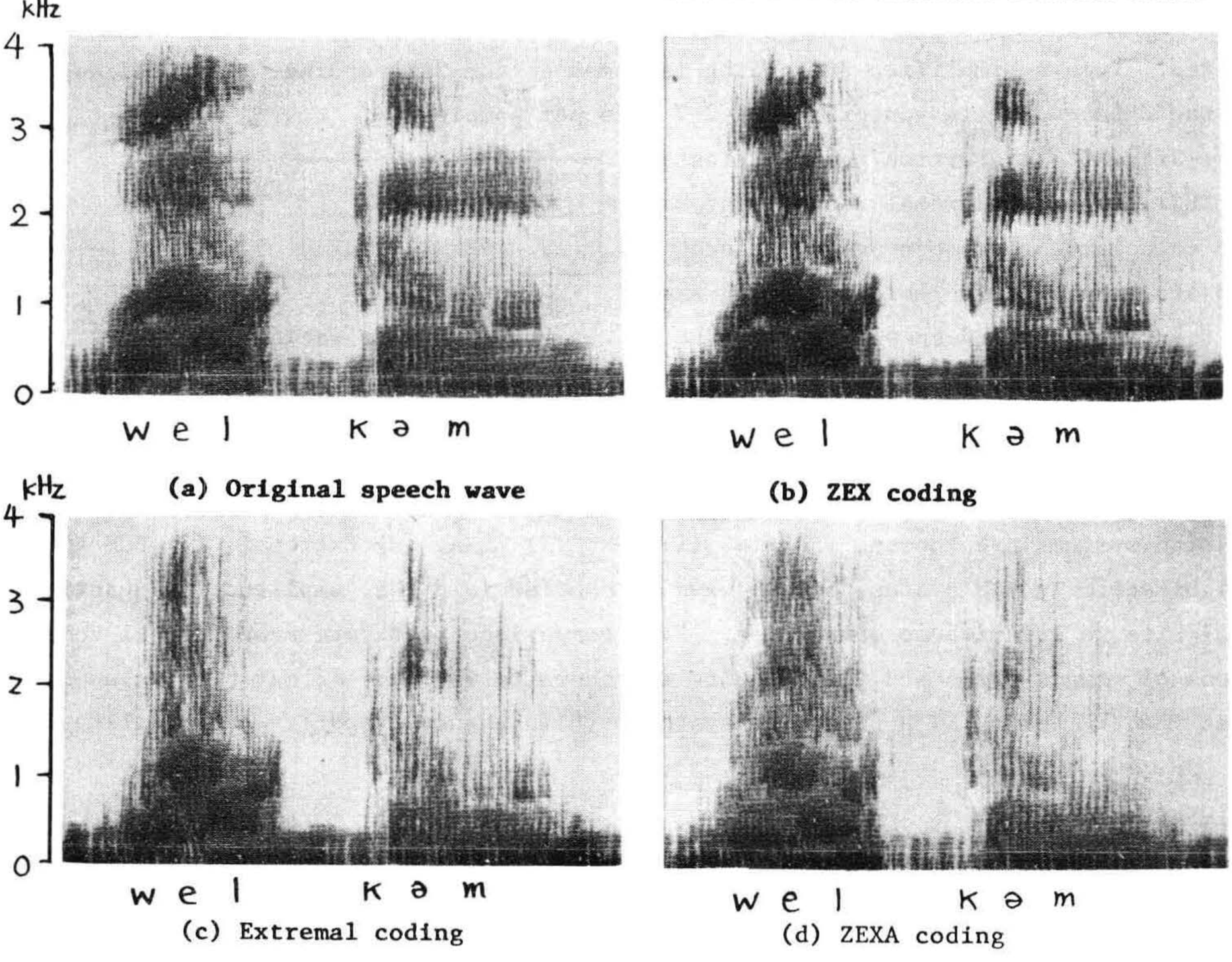


Fig. 4 The average time interval and the number of extremes between zeros

Channel Capacity of ZEX Coding

In a temporal sequence of zero crossing and extreme points, the number of extremes existing between two adjacent zero crossing points depends upon the input speech wave. Statistical analysis of the speech wave of Japanese short sentences are made for the number of extremes and the average time interval between two adjacent zero crossing points. An example of such analysis is shown in Fig. 4. Using these results,



(a) Original speech wave (b) ZEX coding

(c) Extremal coding (d) ZEXA coding

Fig. 5 Sonagrams of the reconstructed speech

the rate of applying equations (1) and (2) for reconstruction of speech is estimated and is approximately

$$\frac{\sum_{i=1}^{19} \left(\frac{1}{i}\right) N_i I_i}{\sum_{i=1}^{19} N_i I_i} = 0.54 \qquad (5)$$

in which I is the average time interval, N is the occurrence and i denotes the number of convex extremes.

Channel capacity of the ZEX coding is also estimated and compared with that of the extremal coding for the case that the sampling frequency is 8 kHz, the time interval is quantized to 8 times of the sampling frequency and the amplitude of extreme is quantized to 7 bits. Results are about 4.8 bits and 3.8 bits per sample respectively. Considering that the occurrence of only one extreme between two adjacent zero crossing points is more than 60 %, there will be a simplified way to reconstruct speech wave, in which only one extreme out of extremes existing between two adjacent zero crossing points is selected and is transmitted. Such a simplified ZEX coding is named as the ZEXA coding. Channel capacity of the ZEXA coding is approximately 2.7 bits per sample.

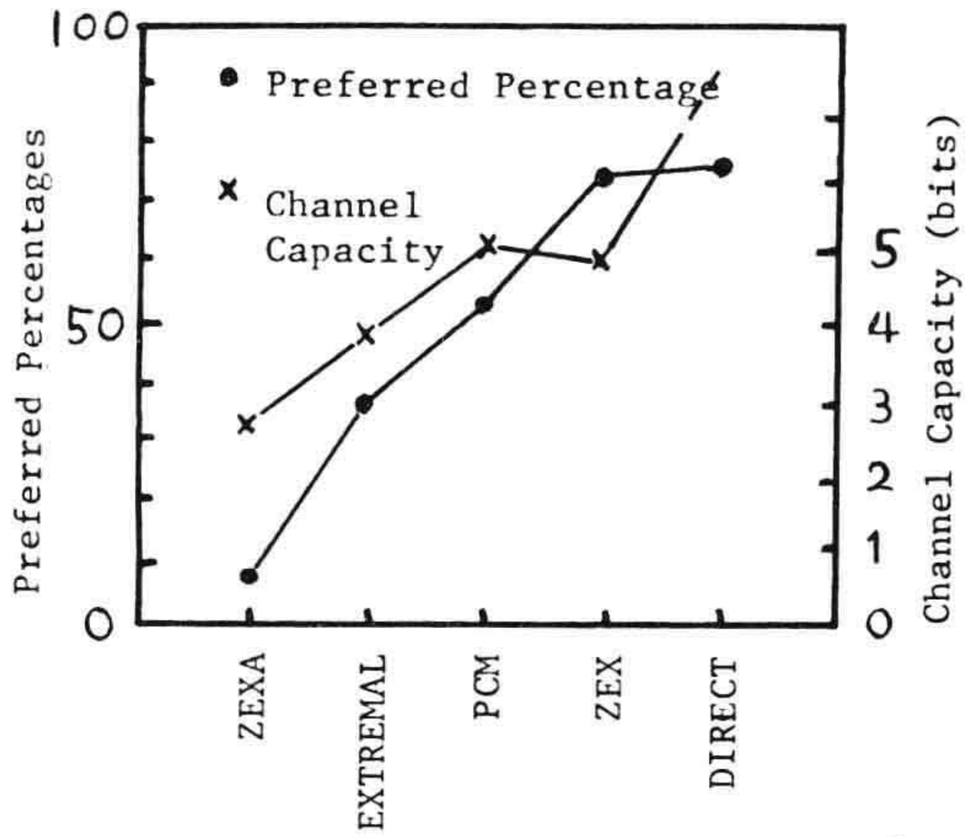


Fig. 6 Pair comparison test results and average channel capacity

Signal to distortion power ratios of ZEX, ZEXA and Extremal coding systems are calculated for Japanese utterrances and are tabulated in Table 1. Sonagrams of the reconstructed speech of these systems are shown in Fig. 5.

ZEX	ZEXA	Extremal
19.0 dB	7.1 dB	11.0 dB

Table 1 Signal to distortion power ratio

Tonal Quality Test of ZEX Coding

To evaluate the tonal quality of the ZEX coding, five kinds of speech transmission systems are chosen, that is, (1) ZEX, (2) ZEXA, (3) Extremal, (4) PCM and (5) Direct. In PCM system, speech wave is sampled in 8 kHz, amplitude is quantized to 7 bits in logarithmic companding. Pair comparison tests are made for all combinations of five systems and test results are shown in Fig. 6. It may be concluded that the ZEX coding is a promising system both in channel capacity and tonal quality.

References:

(1) M. V. Mathews, IRE Trans., IT, 5, p. 129, 1959.

(2) J. C. R. Licklider and I. Pollack, J.A.S.A., 20, p. 42, 1948

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-8

Influence of Bandwidth and Reverberation Time on Preference Evaluation of Speech

G. M. Sessler and J. E. West

Bell Telephone Laboratories, Incorporated

Murray Hill, New Jersey

INTRODUCTION

A subjective study was made to determine the relative importance of bandwidth, center frequency, and reverberation time in communication systems.[1] To implement this study, reverberated test sentences with various bandwidths were presented to subjects in forced-choice paired-comparison tests. The results were evaluated on a digital computer using a factor analytic technique.

PREPARATION OF TEST SENTENCES

The sentence "Should We Chase Those Young Outlaw Cowboys", spoken by an adult male, was recorded in an anechoic chamber. This sentence was then acoustically transmitted and rerecorded in a conference room with variable absorption using reverberation times of 0.1, 0.2 and 0.3 sec. The distance between loudspeaker and microphone in this room was 60 cm. The microphone was at an elevation of 5 cm above a table top. Early arriving echoes, consisting of reflections from the table top and two walls with delays relative to the direct sound of approximately 0.15, 3.5 and 4 msec, were kept constant for all reverberation times. The sentence was also transmitted and rerecorded in an anechoic chamber using the same electroacoustic system as in the conference room. An identical early echo structure was generated by constructing a corner in the anechoic chamber consisting of two walls and a table.

The resulting speech samples with reverberation times between 0 and 0.3 sec were bandpass filtered with lower cutoffs of 250, 350 and 500 Hz, and upper cutoffs of 3, 4.2 and 6 kHz. All nine pass bands that can be obtained with these cutoff frequencies were combined with each reverberation time yielding 36 test sentences or stimuli.

For graphical presentation it is necessary to derive a suitable code for each test sentence. A three-character code was used for this purpose. The first identifier represents the approximate number of half-octave bands contained in the stimulus (see Table 1). The second character identifies the center frequency, where E, L, M, H and V (standing for Extremely low, Low, Medium, High, and Very high) correspond to 0.87, 1.04, 1.23, 1.46, and 1.74 kHz, respectively. Finally, the third identifier gives the reverberation time in 1/10 sec.

SUBJECTIVE TESTING

The stimuli were presented to 11 subjects in forced-choice paired-comparison tests.

½-Octave Bands	Corresponding Frequency Bands (kHz)
5	(0.5 - 3)
6	(0.35 - 3) (0.5 - 4.2)
7	(0.25 - 3) (0.35 - 4.2) (0.5 - 6)
8	(0.25 - 4.2)(0.35 - 6)
9	(0.25 - 6)

Table 1: First Identifier

Each subject judged all 630 pairs that can be formed from the 36 test sentences. The stimuli were reproduced binaurally over earphones at a level of about 70 dB SPL. The subjects were asked to select from each pair the stimulus most efficient for communication systems. Aesthetic quality was not to play a roll in the judgment.

EVALUATION OF SUBJECTIVE DATA

To evaluate the subjective results, the data were analyzed by a factor analytic technique. A preference scale is first derived for each of the n subjects from that subject's paired comparisons data. A computer program, written by J. D. Carroll and programmed by Mrs. J. J. Chang, both of Bell Telephone Laboratories, then constructs a geometric representation of stimuli as points in n-dimensional space, and of individuals as vectors in that same space.[2] The orthogonal projections of stimulus points on individual vectors yield the best approximation, in a least squares sense, to the preference scale values. Each dimension of this space may be thought of as an attribute of the stimuli.

INTERPRETATION OF THE DATA IN THREE-DIMENSIONAL SPACE

We found that dimension 1, or factor 1, accounts for 78 percent of the total variance in the preference data. Twelve percent is contributed by factor 2 and five percent by factor 3, while the remaining eight factors accounted for a total of five percent. Thus the results can be interpreted to a good approximation by the first three factors.

The plane defined by factors 1 and 2 is shown in Fig. 1. Examining the stimuli with respect to factor 1, we find that the first identifier increases in general as we go from left to right. Thus factor 1 corresponds to a first approximation to bandwidth. We know from the individual subject vectors (not shown in the Figure) that preference also increases from left to right along factor 1.

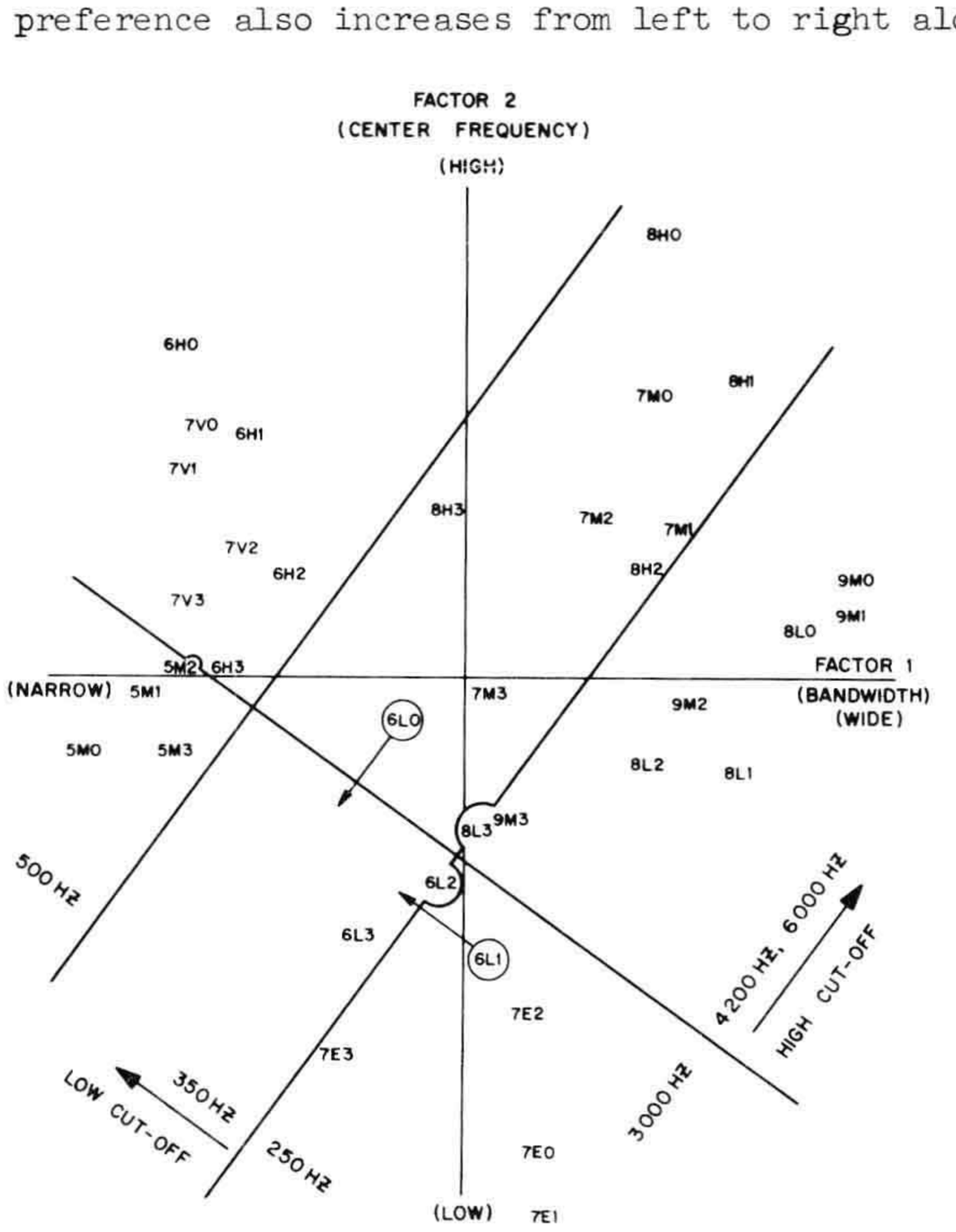


Fig. 1: Plane defined by Factors 1 and 2.

Bandwidth alone, however, does not completely account for factor 1 or preference. Reverberation time seems to be also important for assessing subjective quality. For stimuli with narrow bandwidth (5 and 6 half octaves), it can be seen that some reverberation is desired. In some cases 0.2 sec is preferable to 0 or 0.1 sec. However, if the bandwidth is large, a significant increase in preference can be achieved by decreasing the reverberation time. Since preference depends also on bandwidth, it appears that, within limits, excessive reverberation can be compensated by an increase in bandwidth of the transmitted signal. Examples for this are the pairs (7M0-9M2) and (6L1-8L3) which have about the same preference but different bandwidth and reverberation times. A quantitative formulation of the relation between preference, bandwidth and reverberation time will be given below.

If Fig. 1 is examined with respect to factor 2, it can be seen that stimuli with high center frequencies are located near the top, while stimuli with low center frequencies are located near the bottom. Thus factor 2 appears to correspond to center frequency.

Taking both factors into consideration, the stimuli can be subdivided into 6 groups by 3 lines at angles of approximately 45° to the axes. The two lines with a positive slope with respect to factor 1 separate the stimuli into three regions corresponding to lower cutoff frequencies of 250, 350, and 500 Hz. The line with a negative slope with respect to factor 1 separates the stimuli into two regions with upper cutoff frequencies of 3 kHz and 4.2, 6 kHz, respectively. There are only two stimuli that do not fall into the appropriate areas. It should be possible to draw a fourth line separating the high cutoff frequencies of 4.2 and 6 kHz; however, due to the early echo arriving at 0.15 msec, which gives a spectral zero at 3.3 kHz, stimuli with high cutoff frequencies could possibly not be clearly distinguished.

Figure 2 shows a plot of factors 1 and 3. Recalling that the last identifier corresponds to reverberation in 1/10 sec, we see that stimuli with zero reverberation time are located near the bottom and those with 0.3 sec are located near the top. Thus, factor 3 corresponds to a first approximation to reverberation time. Lines can be drawn to separate the stimuli with different reverberation times. These lines, however, are not of constant angle to the axes which demonstrates that reverberation time is not a linear function of factors 1 and 3. The lines converge with increasing negative values of factor 1 (decreasing bandwidth). This as well as the fact that many stimuli with bandwidth of 3 octaves or less are displaced (this applies to stimuli 5M0 to 5M3 and 6H0 to 6H3 which are therefore omitted in Fig. 2) shows that reverberation time cannot be easily discriminated if the bandwidth is small.

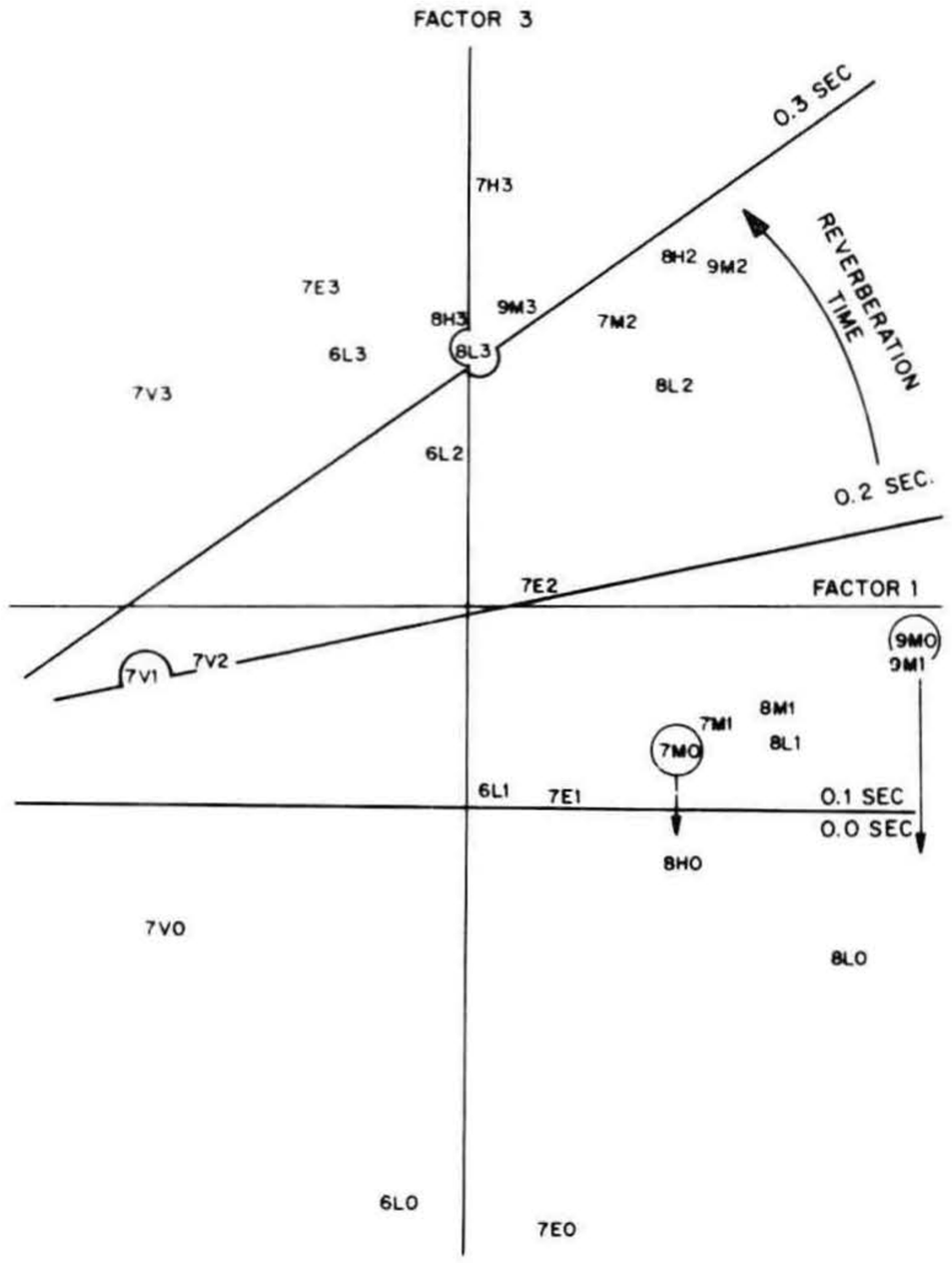


Fig. 2: Plane defined by Factors 1 and 3.

RELATION BETWEEN PREFERENCE AND OBJECTIVE PARAMETERS

Because factor 1 accounts for most of the variance and since it agrees with the direction of the "average" subject vector, we interpret this factor approximately as subjective preference. Apart from bandwidth, however, other parameters seem to contribute to this factor. We derive therefore an analytic expression F′,

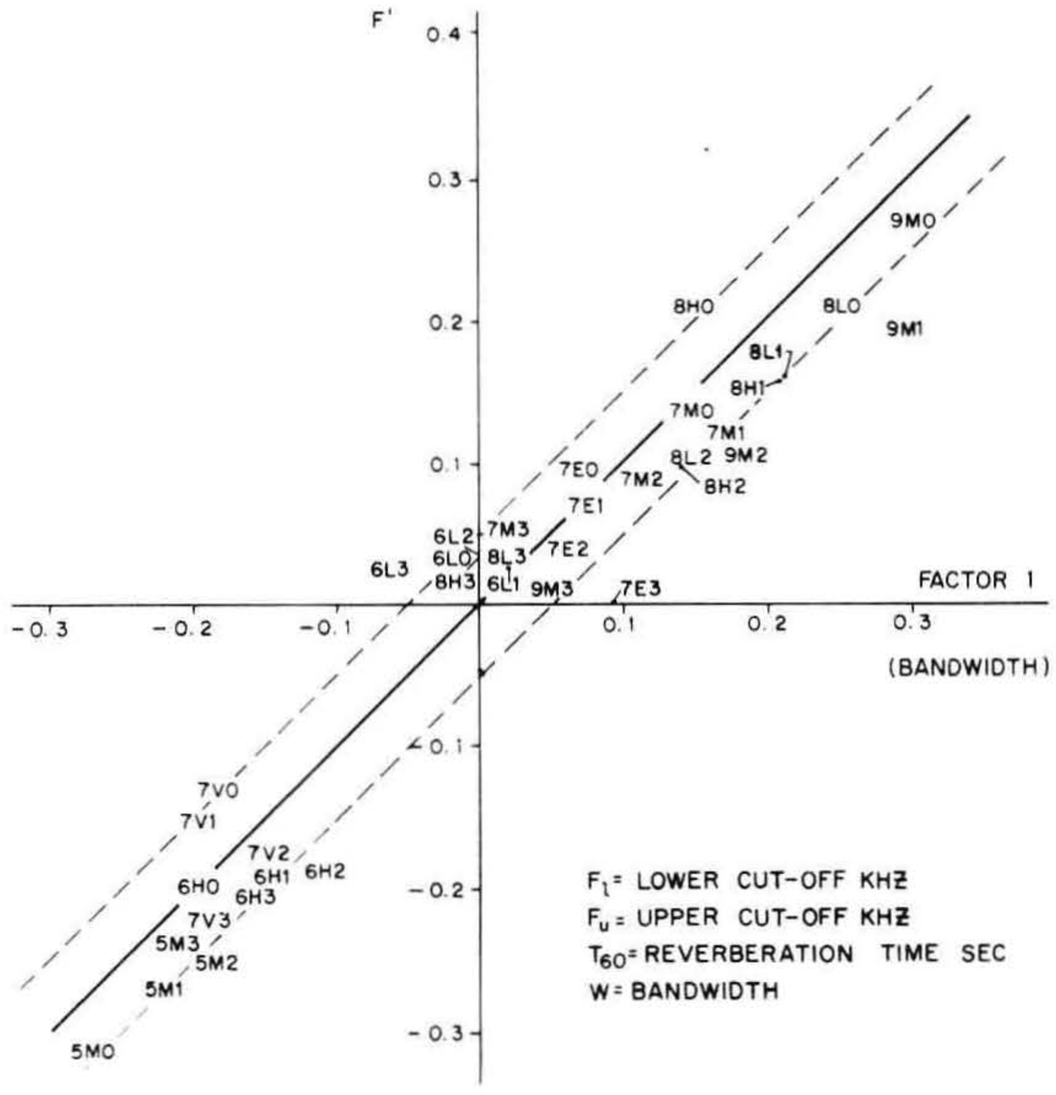


Fig. 3: Plot of F′ computed with Eq. (1). Solid line described by F′ = Factor 1.

$$F' = 0.28 - T_{60}(0.57W+0.5T_{60}-1.9) - 6.6(f_l-0.25)^2 - 0.015(6.5-f_u)^2 \qquad (1)$$

so that F′ is the best possible representation of factor 1 for all 36 stimuli in a least-square sense. As can be seen from Eq. (1), F′ is a function of lower cutoff frequency f_ℓ (in kHz), upper cutoff frequency f_u (in kHz), reverberation time T_{60} (in sec) and bandwidth W (in octaves). In Fig. 3, F′ is plotted against factor 1; the Figure shows that the data are close to the 45° line as would be expected. The standard deviation for the data is 0.04.

A study of Eq. (1) reveals that for small W (3 octaves) the sum in the first parentheses is approximately equal to zero; thus T_{60} has little influence on preference in this case. However, for large bandwidth W (5 octaves), this sum is positive and an increase in T_{60} results in a decrease in preference. Furthermore, the terms with f_ℓ and f_u in Eq. (1) predict an increase of preference with bandwidth. Thus, Eq. (1) reflects the properties of the subjective data.

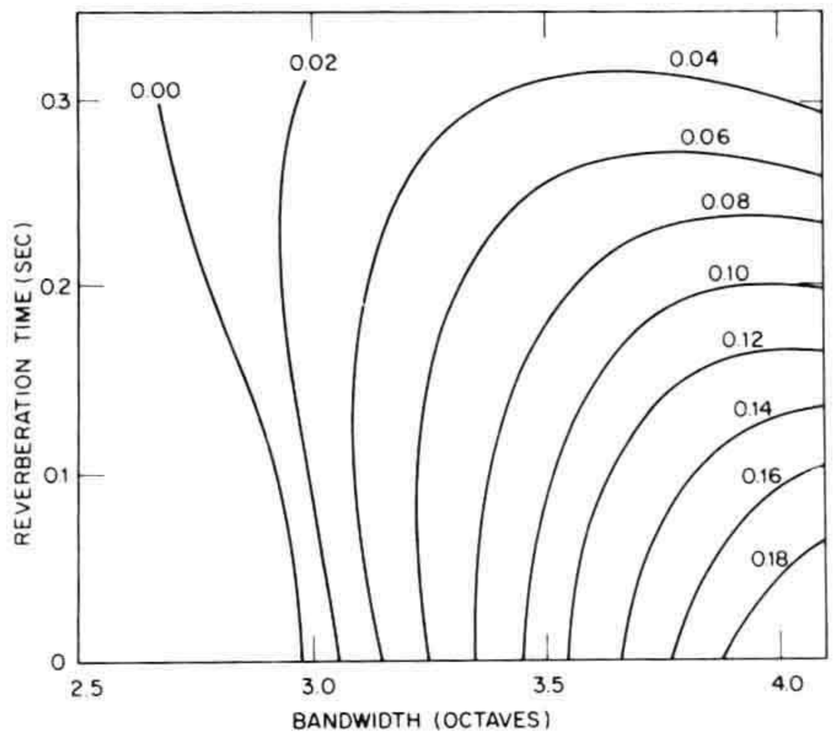


Fig. 4: Isopreference contours in the reverberation time vs bandwidth plane.

The relation between subjective preference, bandwidth, and reverberation time is demonstrated in Fig. 4. The isopreference curves are computed with Eq. (1) for f_ℓ = 0.35 kHz. The figure exhibits the following properties of Eq. (1): (1) for small bandwidth, reverberation time has little effect on subjective preference; (2) throughout the bandwidth range under consideration, stimuli with reverberation times 0 and 0.1 sec show approximately equal subjective preference; and (3) for large bandwidth, preference increases with increasing bandwidth and decreasing reverberation time (except for T_{60} = 0 sec). For example, a sample with T_{60} = 0.1 sec, W = 3.25 octaves is subjectively of equal preference as a sample with T_{60} = 0.25 sec, W = 3.75 octaves. Similar results are obtained for f_ℓ = 0.25 and 0.5 kHz.

SUMMARY AND CONCLUSIONS

Using subjective data, a qualitative relation between preference and objective parameters of speech has been derived. One important aspect of this relation is that, within limits, a change in bandwidth is subjectively equivalent to a corresponding change reverberation time. Such a reciprocal relation can be of importance in future communication systems where it may be easier to increase bandwidth than to suppress excessive reverberation.

More generally, the results presented in this paper show that with the factor analytic technique, the computer can generate from preference data meaningful geometrical representations. It is noteworthy how well these plots can be interpreted in terms of physical parameters. This interpretation may be facilitated by use of one-line computer displays.[3] Thus, this method provides a useful tool for evaluating subjective data.

ACKNOWLEDGMENTS

We are indebted to Dr. J. L. Flanagan and Dr. M. R. Schroeder for inspiring discussions on the interpretation of the results; to Dr. J. D. Carroll and Mrs. J. J. Chang for assistance in applying the factor analysis program to the experimental data; to Miss B. J. McDermott for discussions on the multidimensional scaling method; and to Miss I. Esteves for her help in the experimental phases of the study.

REFERENCES

1. G. M. Sessler and J. E. West, J. Acoust. Soc. Am. 42, 1199 (1967).
2. J. D. Carroll, "Nonparametric multidimensional analysis of paired comparisons data." Unpublished.
3. H. S. Magnuski, G. M. Sessler and J. E. West, J. Acoust. Soc. Am. 42, 1199 (1967).

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-9

Recherches récentes dans le domaine de l'Acoustique et du Traitement des signaux de parole menées au C.N.E.T. (première communication)

Gilbert FERRIEU

Ingénieur des Télécommunications au
Centre National d'Etudes des Télécommunications de LANNION
(Département Etudes et Techniques d'Acoustique) (FRANCE)

Les recherches entreprises par le C.N.E.T., dans le domaine de l'Acoustique de la parole, constituent en fait des applications du Vocoder étudié dans nos laboratoires.

1 - LE VOCODER C.N.E.T.

Sa structure de base est celle, très classique, des Vocoders à canaux. L'analyseur comprend deux parties :

un analyseur de spectre, tout d'abord, constitué de 18 canaux de structure identique ; chaque canal comprend un filtre passe-bande suivi d'un détecteur d'énergie, d'un filtre passe-bas et d'un organe de quantification. Les 18 filtres passe-bande des canaux du Vocoder couvrent, sans lacune, la bande de fréquence de 200 à 8000 Hz environ. Leur espacement est rigoureusement logarithmique. Ces filtres réalisés à l'aide de circuits actifs sont des filtres du 4ème ordre, constitués de deux circuits résonants associés en série. La bande passante à 6 dB de chaque filtre est sensiblement égale au 5ème de la fréquence centrale du filtre ; l'affaiblissement, d'autre part, est de l'ordre de 40 dB à l'octave inférieure ou supérieure de la fréquence centrale. Les détecteurs de niveau d'énergie sont constitués de redresseurs à double alternance. Les filtres passe-bas sont des filtres du 3ème ordre, du type Butterworth ayant une fréquence de coupure de 25 Hz. La tension délivrée par les filtres passe-bas est codée ensuite suivant une échelle logarithmique sur une dynamique de 64 dB par échelons de 4 dB, ce qui correspond à un codage à 4 moments binaires.

La deuxième partie de l'analyseur est constituée par le détecteur de mélodie qui a un double rôle :

- d'une part, distinguer entre sons vocalisés et sons non vocalisés ;
- d'autre part, mesurer de façon précise la période inverse de la fréquence fondamentale de vibration des cordes vocales.

Le principe de ce détecteur de mélodie est le suivant : le signal de parole, filtré par un filtre passe-bas, de fréquence de coupure voisine de 500 Hz, est appliqué à l'entrée d'un détecteur de crête qui fournit une ou deux impulsions par période de mélodie, dans le cas d'un son vocalisé, et des impulsions distribuées aléatoirement dans le temps, dans le cas d'un son non vocalisé.

Un algorithme câblé détermine s'il y a une récurrence dans la suite des impulsions délivrées par le détecteur de crête et mesure éventuellement leur période de récurrence dans le cas des sons vocalisés.

Le synthétiseur du Vocoder comprend lui aussi deux parties.

Tout d'abord, l'ensemble de synthèse sonore, proprement dit, constitué de 18 canaux de stucture semblable. Chaque canal comprend un modulateur qui est en réalité une ligne d'affaiblissement à commande numérique et un filtre passe-bande qui est identique au filtre passe-bande correspondant de la partie analyse du Vocoder. Les signaux de sortie des filtres passe-bande sont additionnés pour constituer la parole synthétique délivrée par le Vocoder. La seconde partie du synthétiseur est le générateur d'excitation artificielle qui attaque l'entrée des 18 modulateurs. Initialement, ce générateur d'excitation artificielle délivrait des impulsions plus fines, donc à spectre très plat, dans toute l'étendue de la bande de fréquence de parole, récurrentes à la fréquence de mélodie dans le cas des sons vocalisés, distribuées aléatoirement dans le temps, pour les autres sons. Une amélioration sensible de la qualité de la parole synthétisée par le Vocoder a été obtenue en remplaçant cette suite d'impulsions par des signaux binaires formant des séquences pseudo-aléatoires de longueur maxima, ayant la même fréquence fondamentale. Ainsi les répartitions spectrales de puissance de ces deux signaux sont les mêmes à une différence de niveau constante près, intéressante d'ailleurs ; il en va tout autrement de la distribution des phrases.

Dans le cas du signal d'excitation à séquence maximale, l'énergie est répartie de façon uniforme dans le temps, quelque soit

.../...

la période de répétition de cette séquence maximale alors que dans le cas des impulsions isolées, l'énergie est évidemment très localisée dans le temps au moment de chaque impulsion. L'utilisation de ce signal d'excitation à une séquence maximale permet d'éviter le caractère un peu heurté de la parole synthétisée généralement par les Vocoders et dû principalement aux transitions entre sons vocalisés et sons non vocalisés.

2 - APPLICATIONS DES VOCODERS A L'INFORMATIQUE

2-1 Le principe des unités à réponse verbale utilisant un calculateur connecté à un Vocoder est bien connu. Je ne citerai, par exemple, que l'unité à réponse verbale fabriquée par la firme I.B.M., sous le nom de I.B.M. 7772. Nous avons au C.N.E.T., abordé ce problème afin de résoudre élégamment le problème de la télétaxation, dans les systèmes de commutation automatique, à gestion centralisée. Lorsque une abonné au téléphone veut, par exemple, connaître l'état de son compteur, ou veut connaître la durée de la dernière communication qu'il vient d'échanger, il peut numéroter le numéro du calculateur de la taxation et celui-ci lui répondra verbalement Pour cette application particulière, le vocabulaire à mettre en mémoire dans le calculateur est assez faible puisqu'il consiste en une ou deux phrases-type et ensuite, la suite des syllabes ou des mots nécessaires pour énoncer un nombre, par exemple, compris entre 0 et 1 million.

La mise au point de cette machine parlante nous a amenés à résoudre deux problèmes de programmation.

- Le premier problème, peut être un petit peu particulier à la langue française qui répugne à épeler chaque chiffre d'un nombre, consiste à savoir comment prononcer une suite de chiffres écrits. Par exemple, les nombres 2 et 5 inscrits dans cet ordre se prononceront "vingt-cinq". Les nombres 2, 5, 0 inscrits dans cet ordre se prononceront "deux cent cinquante".

- Le deuxième problème, beaucoup plus acoustique, d'ailleurs, concerne les corrections de rythme ou d'intonation à effectuer sur la parole synthétisée. Par exemple, les nombres 3850 et 1853 s'expriment au moyen des mêmes mots, mais ces mêmes mots, dans les deux cas, ne sont pas prononcés avec les mêmes intonations, les mêmes fréquences de mélodie et le même rythme. Il est donc nécessaire de prévoir dans la machine parlante, un programme qui

.../...

corrige automatiquement cette intonation, ce rythme en fonction du nombre à prononcer. Ce programme, établi d'une façon très empirique et encore imparfaite, agit évidemment directement sur des éléments binaires d'information du code Vocoder. Signalons, en fait, qu'un Vocoder numérique, même s'il n'est pas couplé à un calculateur permet de créer des effets acoustiques assez spéciaux. Il est possible, par exemple, de ralentir ou d'accélérer à volonté le débit de la parole sans en changer le timbre simplement en enregistrant sur un magnétophone numérique la suite des éléments binaires sortant de l'analyseur et en relisant ensuite cette bande magnétique à une vitesse différente, les informations binaires commandant cette fois le synthétiseur.

2-2 Des méthodes de reconnaissance automatique de parole mettant en oeuvre le Vocoder associé à un calculateur ont fait l'objet d'investigations. Si la mise au point de ces méthodes s'est avérée ardue, certaines semblent néanmoins riches de possibilités pour des problèmes limités de reconnaissance.

3 - APPLICATIONS AU TRAITEMENT DE SIGNAUX

3-1 Correction des distorsions provoquées par l'Hélium

3-2 La Téléphonométrie

x x x

x x

x

Des recherches sont également menées en acoustique générale et électroacoustique.

En particulier, une méthode originale d'étude des haut-parleurs a été étudiée. La réponse transitoire du transducteur est codée puis traitée en temps différé par un calculateur. Ce traitement a pour but de mettre en évidence les réponses amplitude-fréquence et phase fréquence du dispositif. Il doit permettre de relier entre eux certains critères de qualité et en faire apparaître de nouveaux. Un des intérêts de la méthode est de ne pas nécessiter impérativement l'emploi d'une salle sourde.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-10

Universal Phonetics as an Explanatory Theory

Victoria A. Fromkin
Assistant Professor of Linguistics
University of California, Los Angeles

A Theory of Universal Phonetics should provide a framework by which we can describe the sounds of all human languages, and permit a comparison of languages according to a common set of references. It should explain, as well as describe, which sounds and sound sequences can occur and which cannot, and which segments and sequences are most likely to be found in any language and why. It should define the necessary intersecting classes of sounds which are required to make the most general and revealing statements about the phonologies of all languages. Since phonetics is the science of speech sounds, a phonetic theory is intrinsically connected to a theory of human language, i.e. linguistic theory. To speak or understand the utterances in a language implies the existence of rules which enable one to pair certain sounds with certain corresponding meanings. Thus, a theory of Universal Phonetics must allow for an unambiguous representation of these sounds.

These are some of the goals which we seek, and some of the criteria by which we can judge the validity of a proposed Universal Phonetic Theory. They are not new goals. There are many old questions still unanswered. If any headway has been made toward

achieving these goals, we are indebted to all those in the past who, intrigued with man's unique ability to use language, persistently attempted to achieve some understanding of this enormously complex, always perplexing phenomenon.

It is reported that when someone once asked Newton how he was able to make his great discoveries, he replied: 'It is because I stand on the backs of giants.' If modern phoneticians are making progress toward establishing a theory of Universal Phonetics, it is because they, too, stand on the backs of giants.

One notion that has persisted through history is that despite the semi-continuous nature of the acoustic speech signal, human utterances can be quantized into discrete units. The modern instruments now available for phonetic research reveal the analog nature of the signal and the blurring or total obliteration of the more abstract units. But 3000 years ago the Hindus did not need a speech spectrograph to enable them to understand the difference between speech, and knowledge of speech, or language. According to a Hindu myth, 'in response to an appeal made by the gods, the god Indra attempted for the first time to break up speech into its component elements and from thenceforward it could be regarded as language.'

Plato, in the Cratylus Dialogue, associates speech with Hermes, the messenger who is in continuous motion. However, he has Socrates ask: 'How then can that which is never in the same state be anything?' Socrates then proposes to 'examine all these matters with scientific ability, and to take it (i.e. speech) to pieces.'

What is obvious today, and what was obvious to the ancients, is that there can be no Universal Phonetic Theory which excludes the notion of discrete units, which units must be of a higher order of abstraction than can be found in the raw data of speech perfor-

mance.

A second assumption which has persisted is that the range of these discrete units is limited. The finiteness and the universality of the human vocal organs gives rise to this notion which stimulated the attempts to construct universal alphabets, such as the one proposed by Cave Beck in 1657. He described it as an attempt to provide 'The Universal Character by which all the Nations of the World may understand one another's Conceptions, Reading out of one Common writing their own Mother Tongues.' In order to use such an alphabet to describe all languages, the primitive elements must have some intrinsic content which permits a non-arbitrary description. Otherwise it would be impossible to compare the sounds of one language with those of another. Furthermore, the nature of the elements which are chosen determine the adequacy of the classes which they define.

Again there was historical recognition of this fact. The Hindus classified the sounds of Sanskrit by a fixed set of articulatory parameters, for example. Many of the categories they used to define the classes of speech sounds correspond to the features we use today.

While these basic concepts have their origin thousands of years ago, it is clear that phonetic theory has been refined, and equally clear that it stands in need of further refinement.

There is as yet no agreement as to the set of features which should comprise the set of primitive elements of the theory. The set of Distinctive Features originally put forth (Jakobson, Fant, and Halle, 1955) proved to be inadequate. More features were added. Others were deleted. The physical correlates were changed. The desire to establish a set of universal constraints on the sound systems of all languages, i.e. on feature combinations and feature

sequences, and to formalize our intuitive notions of'natural 'classes of sounds and 'expected' phonological rules has led to a resurrection of the early Prague school notion of markedness and unmarkedness, the unmarked specifications being the 'natural' or 'expected ones. To this end the new instruments and techniques for phonetic research can provide insights and explanations, and can test hypotheses. A reawakened interest in physiological parameters has resulted from the fact that acoustic analysis of speech failed to provide the knowledge needed to explain universal constraints. If, as is suggested, taking account of the instrinsic content of features leads to more natural classes, it is evident that the 'intrinsic content' should indeed be intrinsic. When features such as 'sonorant' are defined as 'segments in which spontaneous voicing is possible', this is a statement which is falsifiable in the laboratory, using techniques such as glottography. If the unmarked member is viewed as less complex or more normal than the marked one, only by experimentation can we discover what is the less complex state. The technique of electromyography provides information on which muscles are used to produce different speech sounds, the sequencing of their actions, the timing of the events. We can thus substitute quantitative measures of 'complexity' for vague notions used to date.

Bacon once stated that 'the experiment is a question asked of nature to make her speak.' In this sense work in experimental phonetics can help find answers to questions relating to a Theory of Universal Phonetics.

The research reported on in this paper was supported in part by the National Institutes of Health, U.S. Dept. of Health, Education, and Welfare, the National Science Foundation, and the Office of Naval Research.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-11

Speech wave, articulation and their linguistic function

Bożena Wierzchowska and Józef Wierzchowski

Warszawa - Poland

Maria Curie-Skłodowska University in Lublin

It seems valid to determine the structure of the language taking simultaneously into consideration the form and the meaning. The commutation test used by the Prague School of linguistics, i. e. in the form where the meaning is taken into account in oathegories of distinotiveness only /the meaning of the form A is not the same as the meaning of the form B, but both forms, A and B, can be polysemic/, enables us to establish the inventory of phonemes of a given language and also to distinguish the word forms of this language.

It seems valid to extend the usually applied investigations of the realizations of phonemes where the phonemes are treated as the components of the phoneme-inventory of a given language, and also the investigations of the realizations of phonemes where their environment is taken into consideration /cf. also the postulate of K. L. Pike [1] who proposed to investigate the phonemes taking into consideration the grammatical facts/, and undertake also the investigations of the realizations of phonemes treated as components of various word-forms of a given language.

The realizations of the same phoneme can be different in various word-forms, even if the immediate phonetic environment and the conditions in the rhythm and stress are the same. These differences in the realizations of phonemes can be manifested both in the range of free variation and in the different frequency of various allophones which can be observed in the utterances of the same person as well as in the utterances of a variety of persons. These differences in the realizations of the phonemes can be established statistically. For instance, the Polish word-form d y r e k t o r /»director«/ is often realized as d ə r e k t o r, d e r e k t o r; the realizations of the phoneme y are in this word-form frequently variants of the sound between y and e. In colloquial speech the realizations reach the positions of the main variant of the Polish phoneme e : d e r e k t o r. But the phoneme y in the word-form d y r y g e n t /»leader of an orchestra«/ has a smaller number of variants. This word-form is usually uttered as d y r y g e n t.

It seems also valid to connect the realizations of phonemes not only with the word-form, usually polysemic, but with the semantically uniform unit - the word. A method which enables us to establish semantically uniform word units in the word-forms of a given language, similar to the transformational method used in the syntactical analysis, is proposed by J. Wierzchowski [2].

There are differences in realizations of phonemes which are sometimes clearly associated with definite meanings of word-forms, i. e. with the words which can be distinguished in the word-forms. To illustrate, the variations in realizations of the phoneme ǫ in the Polish word-form w ǫ š /wąż/ are statistically different when the word w ǫ š /»coil«/ is uttered, and different when the word v ǫ š /»snake«/ is uttered. These differences can be

observed both in utteranoes of the same person as in utterances of any group of persons.

The permanent progress in acoustic and articulatory /genetic/ investigations of speech /this progress in articulatory investigations is connected first at all with the use of cine-radiography/ has made it possible to realize the claim of O. Jespersen who supposed [3] that the classification of speech sounds should be made taking simultaneously into consideration their acoustic and articulatory aspects. There can be observed the permanent correspondence between the height of F_2 as it appears on the Sonagramms of various speech sounds and the position of the tongue in the articulation of these sounds [4]. If during the articulation of a speech sound the blade of the tongue is raised forward and upward, as in Polish sounds i, y, ś, ź, ć, ṕ, ḃ, ḱ, ẋ etc., in Russian sounds i, t', d', s', z' etc., German sounds i, ẋ, English sounds i, j, the F_2 raises to approx. 2300 Hz - 2500 Hz. When the position of the tongue is level, as in Polish sounds a, t, d, s, z, š, ž etc., Russian sounds a, t, d, s, z etc., German sounds a, s, z, t, d etc., English sounds a, ʌ, θ, ð, s, z, t, d etc., the F_2 becomes lower, and it appears in the range from approx. 1200 Hz to 1700 Hz. Very low position of F_2, in the range from approx. 700 Hz to 900 Hz, corresponds to the upward retraction of the blade of the tongue, as in Polish, Russian, German, English sounds u, x, k etc. When the blade of the tongue makes a closure with the lowered soft palate and so the nasal cavity is involved into articulation, an interruption in the registration of F_2 can be observed which is caused by elimination of the role of the tongue-position in so uttered sound.

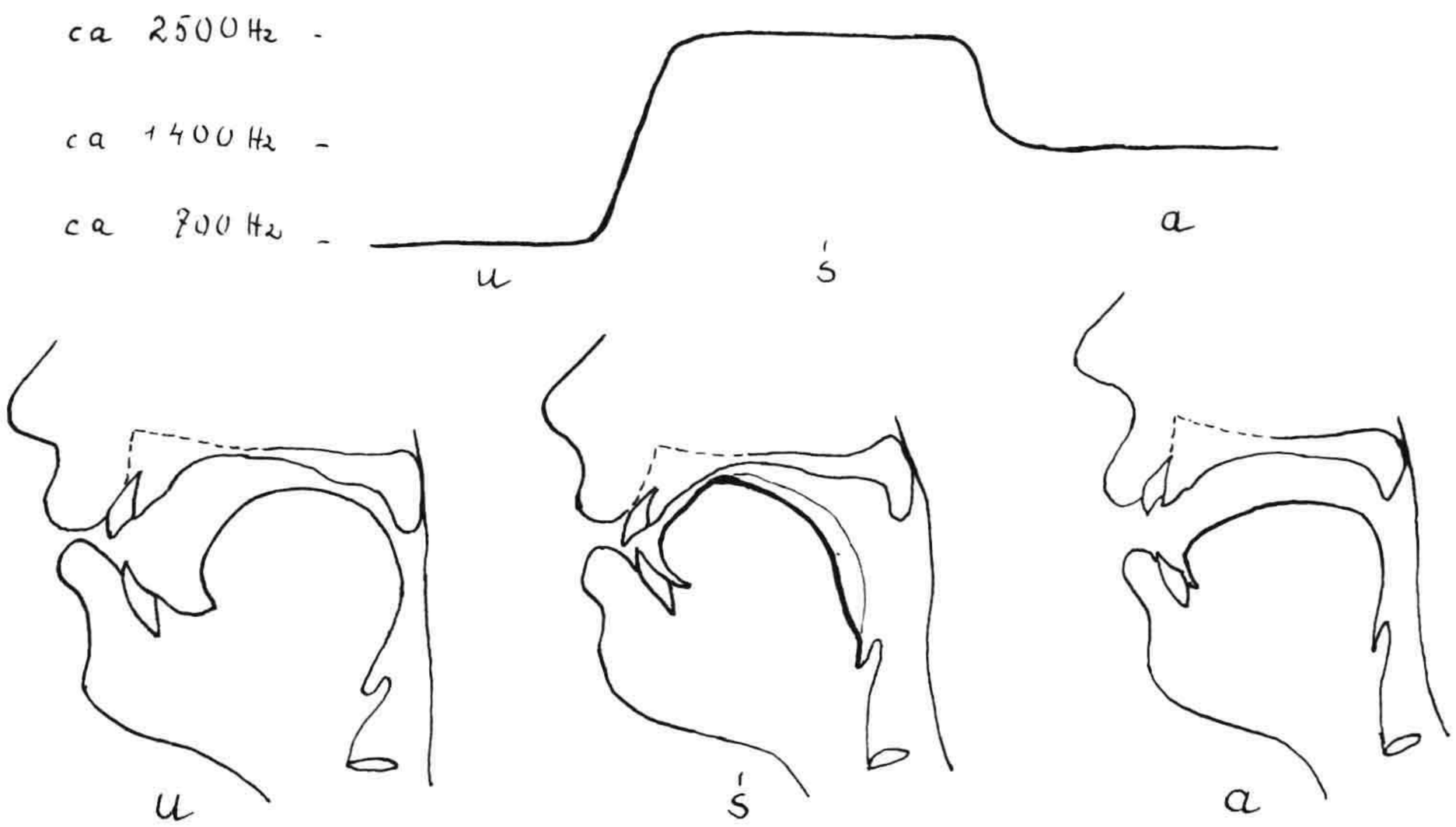


Fig. 1. The height of F_2 in the sound sequence u ś a and the corresponding positions of the tongue.

Referenoes.

1 K. L. Pike, Grammatical Prerequisites to Phonemic Analysis, Word, 3, 1947, pp. 155 - 172.

2 J. Wierzchowski, Le mot et la structure phonologique et sub-phonématique, Study of Sounds, XII, Tokyo 1966, pp. 102 - 114.

3 O. Jespersen, Akustisch oder genetisch? In : Phonetische Grundfragen, Leipzig-Berlin 1904, pp. 65 - 104.

4 B. Wierzchowska, A Cine-radiographic Examination of Some Polish Phonemes, Phonetica, 13, 1965, p. 114. The Polish nasal sounds are discussed in detail by this author /taking into consideration simultaneously their articulatory and acoustic aspect/ in her : Analiza eksperymentalno-fonetyczna polskich dźwięków nosowych, Lublin 1966 /in Polish/.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-12

ON LEARNING LINGUISTIC CONCEPTS

Ernest Colton

Consultant

3924-¼ Candlewood Avenue

Lakewood, California 90712

If nature abandons its unsuccessful experiments and management expands its unsuccessful experiments, then which will survive?

Anonymous

An important problem in acoustical linguistics is the origination, coding, sending, receiving, decoding, and recognition of speech symbols or sounds. As soon as one deals with subjects about which one's knowledge is inherently incomplete one gets stuck with difficulties generally masked under the heading of "subjective variables". Obviously, it is desirable to clearly separate such unobservable variables from observable ones, such as physical sound patterns.

It has been shown[1-4] that it is feasible to define the observable and unobservable parameters of communicating systems under very general conditions (implying that there can be no exceptions to this theory under all such conditions). This result is based on the fact that the process of communicating produces a meaning change in a system consisting of communicators and their interconnecting communication channels. One must thus be interested in the change of such variables as "semantic meaning" (synchronization patterns)[1,2], "lexical meaning"[3,4], "significance"[1]. Since these variables can change with time one is not only interested in how these variables change but in how they behave in time. Thus, the problem could be stated formally in terms of a Pfaffian expression.[1,4] (And it should be emphasized that one requires the existence of a solution (i.e. the integrability of the Pfaffian expression) on physical grounds and not vice versa!)

The above formalism resulted in such interesting concepts as "meaning death"[3] and "absolute" and "empirical scales of lexical meaning"[4]. It brought out rather forcefully that lexical meaning is an intensive (not additive for different systems) variable. It is an observable property whose value does not depend on its past history but is defined only from its instantaneous state while its conjugate variable, semantic meaning, has more or less opposite characteristics. An element of semantic meaning turns out to be roughly equivalent to an element of significance per unit absolute lexical meaning. It should be noted that the concept of lexical meaning is only useful because a communication or meaning system is (almost always) not fully known and one wishes to derive some valid conclusions from one's incomplete knowledge.

It was also noted that a communication channel can only carry information (and add noise to it) and switch its connections to/from receivers, objects, senders.[3] Thus, it was deduced that lexical meaning can only be a function of channel information and channel switching. The latter channel property turned out to be a basic intensive variable of speech communicating systems. It is the conjugate variable to information. Switching is defined by the total channel switches (during an appropriate observation time) divided by the total number of "brains" (weighted by categories if required) within the

system.

A useful conceptual model is that of a meaning system consisting of a very large number of "brains" capable of "learning" a "language", "connected" to each other and an environment by various communication channels.[1] This model permitted one to use continuum theory (i.e. sets consisting of a type of serial order) and thus to avoid the statistical trap of dealing with concepts based on variables "defined" by unordered sets (making any crosscorrelation invalid).[5] (It should be noted that a concept defined in continuum theory, e.g. momentum, height, weight, etc., belongs to an ordered set even if its measure is derived statistically from a series of observations.)

Ashby[6] has given a very general description of how systems behave as they are disturbed from an equilibrium condition. Upon applying his results to the above model of a meaning system it becomes immediately obvious that all communication channels cannot be equally accessible to any particular "brain element" of such a system. It should be noted that Ashby's method requires one to consider the above defined meaning system in terms of very many different variables rather than the few general ones discussed above. In the remainder of this paper it should be clear from the context (if not otherwise indicated) as to what variables are used and why.

Consider a brand new meaning system with its "brains" and their communication channels interlinked at random. Obviously, such a system would not be in equilibrium (in Ashby's sense) — by a very high probability. Thus, it would start to communicate. The result of this must be that, quite aside from necessary geographical aggregates (which may in part be due to "inherited" structural locations), various other meaning subsystems will form. These subsystems will be in at least partial isolation from each other (poorly joined systems with temporary independences (constancies across interconnections) in terms of Ashby's description; in limited communicational contact or "unswitched" in terms of this author's approach). These subsystems will develop their own (specialized) sublanguages along with their concommitant synchronization patterns.[1]

Similar considerations will establish that an existing large meaning system must also develop such subsystems as soon as it adjusts back to equilibrium conditions from disturbances.

It is obvious that each of such subsystems must approach its internal meaning equilibrium[3]. In the boundary region between two communicationally adjacent subsystems the military (change control) function[3] becomes of prime importance. Typically, recruitment of additional "brains" and their communication channels occurs in this region until a strong (counter) independent subsystem is encountered.[5] After this the two systems adjust to each other by reducing uncontrolled[3,5] contacts with each other or by eventually merging into a single subsystem. A "brain" well within a particular subsystem (i.e. communicationally far away from its boundary region) thus does not have equal accessibility to brains and communication channels in other subsystems (assuming it has such accessibility within its own subsubsystem).[5]

This accessibility can thus be expressed in terms of a meaning potential function[5]. At the boundary between two communicationally adjacent subsystems this function approaches the same value for each of the two subsystems, while at their communicational centers the values would differ. Since there are many different subsystems there are many accessibility values and, hence, meaning potentials. However, one is usually only interested in the potentials of a limited number of pertinent subsystems.[5] Some of these potentials will be attractive and the rest repellent to a particular brain under specific conditions. (The details of these potential values are a function of the synchronization signals and lexical meaning levels.[5])

It has been previously derived that the "church" (identity preservation) function is intimately connected with meaning death[2,5]. It results in pre-empting for synchronization signals more and more of the communication channels available to a meaning (sub)system. This reduces working contacts with other (sub)systems. In terms of Ashby's description the (sub)system would tend to become unjoined from its environment and, hence, eventually to lose its adaptive capabilities.

From the above discussion it follows that semantic meaning decreases[1] spontaneously, in specified circumstances, simply because a meaning system tends towards an equilibrium. Meaning death[3] is thus usually the natural condition which occurs as a result that equilibrium is more probable (because it is achievable in a larger number of ways) if subsystems develop. This does not imply that it would not occur (at a much slower rate) even if no subsystem would develop. Rather it is to say that meaning death of subsystems will occur first and this usually terminates a particular meaning system long before it would otherwise "die".[5]

Consider now an aged biological brain as a meaning system. Obviously, the language of interest

here is that of the brain. (This is different from considering large social linguistic systems as meaning systems.) Assume this system consists of a large number of nerve cells and their interconnecting communication channels. (The nerve cells are here considered as the individual ''brains'' of the meaning system; it is assumed that they are capable of ''learning'' the brain (i.e. nerve) ''language''.) Some of these channels are connected to internal and external body sensors. One can weight these sensory (and other) communication channels for physical age deterioration.

From the previous discussion it follows that such a meaning system consisting of an aged biological brain must be highly structured. It must consist of subsystems in minimal (or constant synchronizational) communicational contact with each other. These subsystems have developed (sub)languages and synchronization patterns which result in minimal adjustments to environmental disturbances. Unfortunately this tends towards a loss in adaptibility (or learning ability) for the biological brain as a whole. This loss must be much greater than that due to cell losses and other ageing processes.

A little reflection on the above phenomenon indicates that in a young brain the switching variable is at a maximum value and the information variable at a minimum value while in an aged brain (in a familiar environment) the reverse situation exists. It is thus clear that a gerontological dilemma can occur. If the aged brain system is not exposed to new stimulations (which, from a practical point of view, can usually come only through the exterior sensory communication channels) it will sink deeper into meaning death (and eventually suffer from concommitant cell atrophy). If it is stimulated too strongly its equilibrium could become so disturbed that the system ''falls apart''[4] (i.e. the organism ceases to function or dies). Also, in a free environment an aged brain system will tend to avoid exposing itself to disturbances (quite unlike a young brain). Again, this follows directly from the previous discussion. Much work remains to be done on this problem (which will remain even after presbycusis effects are overcome). It is almost certain that speech sounds will have to play a major role here but the how and what are not yet known. One interesting approach is to consider a group of aged brains as a single meaning system.[5]

Consider now a single biological cell as a meaning system. Obviously, here the language of importance is a physicochemical one. The ''brains'' capable of ''learning'' (over a genetically sufficiently long time span) the ''language'' are such 'molecules' as genes. From the previous discussion we must expect that the communication channels within the cell will tend to form subsystems. Since in a cell such subsystems tend to have an observable location,[5] one must conclude that an aged cell should show more and larger subsystems than a young one. (The channel subsystem location does not necessarily coincide with the gene location.) Also, it is clear that such subsystems must eventually result in the biological death of the cell system.

If a cell divides such as to pass on to its ''offsprings'' a substantial portion of its individual language and its sublanguages, then the total number of such cell divisions must be expected to be limited for similar reasons. An exception to this could theoretically occur if a subsystem can develop to where it can overwhelm all countersynchronization signals at its boundaries. The cell would then no longer be functional in its normal way but would very likely undergo rapid growth and mitotic reproduction. It would have become a type of cancer cell.[5] From the previous discussion it is obvious that such growth can be brought under control by countersynchronization signals. Thus theoretical evidence exists that physiocochemical means (i.e. the cell language) could control these types of cancers.

It is thus clear why a ''church''[2] after 'wrecking' its ''religion''[2] and 'ruining' the ''military''[3] will have brought its particular meaning system close to its final destruction.

It has been previously mentioned that learning a major new concept involves a phase change of the meaning system undergoing the learning.[3] From the previous discussion it follows that such a phase change involves, at least for a time, the creation of a new meaning subsystem. This implies, at least temporarily, a change in stability of the system, i.e. the system must become at least partially unstable. From the criteria of intrinsic stability of a system and the previously discussed meaning potentials one can draw some conclusions about the learning process.

It is obvious that the previously discussed meaning potential function can be expressed in terms of powers of meaning potentials. This implies extremum points in the channel accessibility curve (which ''wiggles''). In turn, this implies fluctuations in the values of the variables describing the state of the meaning system. A typical fluctuation applicable to the learning of a single concept is shown in Figure 1. In this figure is shown a single ''isolexical''[4] curve as a function of the switching and information variables (which, in a practical situation may include some weighting factors, e.g. acoustical or optical properties). (Other isolexicals could normally be expected to have different extremum point locations. However, this is irrelevant to the

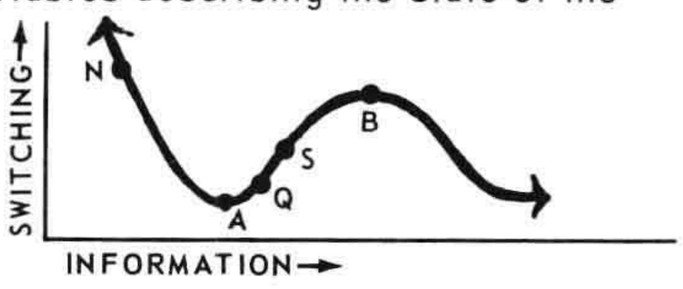


FIGURE 1

present discussion.) It is obvious that the system could not be intrinsically stable between points A and B. A phase transition must occur here.

Consider the effect of "educating" an adult, say an industrial research manager, in one of the, in the U.S.A. at present quite popular, science updating courses. Typically, such an individual supposedly learns to comprehend such major concepts as quantum mechanics (relativistic at that), cybernetics, topology, psychoacoustics, etc. each in a period of about five hours. The background of such a student is usually such that he can be considered to be a complete novice to the field; his mathematical knowledge often consists only in being able to ask his secretary to "run some figures" for him.

It is obvious that exposing such an adult brain meaning system to such an educational condition might move it from point N to point Q. The subsystem mechanism discussed earlier would move it back to point A in a rather short time interval after the end of the course (if it could get past point B it would be a different matter but this is not feasible in such a short period and is not attempted).

So far so good, the student has presumably learned "a little bit". But the aim of these courses is presumably to allow the manager to communicate on a technical level with specialists in the various fields. These people (brain meaning systems) are far to the right of point B. Obviously, no detailed technical communication is possible between them and the above manager. However, the manager is now able to go into the standard managerial act of innuendo and managerial phrases[5] by including some technical words and (freshly learned) questions. But with this inclusion he is able to get past the synchronization defenses of the technical people (meaning systems).[1] But since the manager has basically nothing to contribute at the technical level the result can only be one of equivocation.[5] The damage (disturbance effect) is rather greater than might be expected because such a manager usually controls some key channel in the industrial communication structure (in an organization not all channels carry equal weight and communication possibilities).[5] As could be expected, the system adjusts by changing the synchronization signals, i.e. by subtly changing the technical "jargon".[1] This leaves the brain meaning system (of the manager) worse off than before it was exposed to such courses because it is not able to get far away from either side of point A. Its subsystems will undergo some temporary joining when the proper word inputs occur. But since the synchronization is lacking, this results only in (weak) equilibrium disturbances. (The manager "feels" that he ought to understand the communication but "feels" rather confused and frustrated (which could, in extreme cases, lead to headaches and nausea).[5])

Thus it can be shown that this type of education serves no utilitarian purpose (except for the course organizers); on the contrary, it damages the larger society of which it is a part. However, since they are proliferating one should, perhaps, ask: why? An answer could be readily given in terms of the meaning death of the system, however, it has been more poetically expressed by Veblen[7].

It should be remarked that adult education designed to get (working, technical) people from point S past point B is an entirely different matter. Its results would obviously be quite different than the ones discussed above. (But it could not expect to teach major concepts in a short time and without effort by the students.)

Lastly, it should be remarked that any education (indoctrination) (for young, adult, or aged brain meaning systems) which is not designed to get the student past major isolexical maxima would be a disservice to its students and supporters. However, more knowledge of isolexicals is required before this criterion can be applied to all conditions.

Learning of all linguistic concepts must involve the above mechanisms[1] although the isolexicals would usually not undergo such pronounced fluctuations during the learning of an average (word) concept. It is thus clear that the mapping of isolexicals is an essential tool for the study of the coding of meaning into acoustic speech symbols or sound patterns. Without it one has great difficulties in deciding just where and how one's study of speech sound patterns fits into the communication problem with which this paper was introduced.[5]

REFERENCES

1. E. Colton, On the Contribution of Synchronization Patterns to Semantic Meaning, Paper H22, Fourth International Congress on Acoustics, Copenhagen, August 21-28, 1962. (Read "lexical meaning" in lieu of "significance" in the second paragraph.)
2. E. Colton, Oral portion of the paper of reference 1.
3. E. Colton, On Measuring Lexical Meaning, Paper B62, Fifth International Congress on Acoustics, Liège, September 7-14, 1965. (Add "as data digesters" just after the closing parenthesis in the last sentence of the 4th last paragraph on page 4.)
4. E. Colton, Oral portion of the paper of reference 3.
5. E. Colton, Unpublished notes.
6. R. Ashby, Design for a Brain, Chapman and Hall Ltd., London, 1966.
7. T. Veblen, The Theory of the Leisure Class, New American Library, New York, 1953.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-13

System Theoretical Considerations on Speech Generation, Transmission, and Processing

Syuji Inomata

Principal Scientific Official, Electrotechnical Laboratory
Ministry of International Trade and Industry, Japanese Government

Introduction

In this study, a "general" model of speech hearing and speaking systems is proposed, and several aspects of speech generation, transmission and processing have been discussed by system theoretical considerations.

The most interesting result is the system theoretical justification of the "active"[1] or "analysis-by-synthesis"[2] method of speech analysis and recognition. The method has been thought to be most promissing among various speech recognition methods which have been proposed or exprimented, but it lacks direct and explicit justification, however, has indirect supporting evidences from several aspects of speech perception phenomena.

Also investigated are other related problems such as the speech generating and cocktail party problems, speech transmission between normal and "auditory-handicapped" models and others.

Foundamental Requierments and Assumptions

The "general" model of speech hearing and speaking system has been proposed under the following foundamental requierments and assumptions.

1) Hearing and speaking processes must be taken into account simultaneously into the model, because man pronounces speech sounds as he hears them.

2) The model has two kinds of output, "auditory" and "articulatory". Where articulatory output means the movements of speech generating organs. And only auditory feedback is considered in the discussions.

3) Speech wave is continuous, and the internal state of the system is also assumed to be continuous.

General Model

According to the above requierments and assumptions,and also, to the traditional manner of the definition of a system or automaton, the general model is defined as follows.

$$\left.\begin{aligned} \frac{du}{dt} &= F(u,v) \\ w &= G(u) \\ x &= H(u) \end{aligned}\right\} \qquad (1)$$

Where v is speech wave fed to the model, w, generated speech wave, x, vector articulatory parameter, F, G, H, indicate functions or functionals which characterize the model. It should be noted that speech wave and articulatory parameters are related by the following functional relation, that is,

$$w = I(x) \qquad (2)$$

It should also be noted that, in reality, generated speech wave, w,is directly fed into the ear, that is, w=v holds, and so, the above formulation is rather "ideal" and "hypothetical".

Speech Generation Problem

If the model "speaks" in a silent space or anechonic chamber, the following can be obtained, that is,

$$\frac{du}{dt} = F(u,v) = F(u,w) = F(u,G(u)) = F_o(u) \qquad (3)$$

from the equations (1), (3), we obtain,

$$\left.\begin{aligned} \frac{du}{dt} &= F_0(u) \\ w &= G(u) \\ x &= H(u) \end{aligned}\right\} \qquad (4)$$

We call the above set of equations, "speech generating equation" of the general model. If F_0 is a simple "function",and the first order differential equation can be solved, we can rewrite the above expression as below using the internal state u_0 at t=0, and time t, that is,

$$\left.\begin{aligned} u &= F_{0,0}(u_0, t) \\ w &= G(u) \\ x &= H(u) \end{aligned}\right\} \qquad (5)$$

Speech Recognition Problem

We consider the case where the speech models A_1 and A_2, which have the same structure, are in conversational state, and the model A_1 in speech generating state. As to the model A_1, we obtain the following speech generating equation.

$$\left.\begin{aligned} \frac{du_1}{dt} &= F_0(u_1) \\ w_1 &= G(u_1) \\ x &= H(u_1) \end{aligned}\right\} \qquad (6)$$

As to the model A_2, from the fact that it is in hearing state and from the equation (3), we obtain,

$$\frac{du_2}{dt} = F(u_2, v_2) = F(u_2, w_1) = F(u_2, G(u_1)) \qquad (7)$$

Where, if the internal state of the first model transfers completely to the second model, or in other words, if the Euclidean distance,

$$dist(u_1, u_2) = \left(\sum_{i=1}^{\infty} (u_{i1} - u_{i2})^2 \right)^{\frac{1}{2}} \qquad (8)$$

vanishes for all t, the equation (7) becomes

$$\frac{du_2}{dt} = F(u_2, G_r(u_1)) = F_0(u_1) \tag{9}$$

And as to the second model we obtain the following.

$$\frac{du_1}{dt} = F_0(u_1) \tag{10}$$

The above is the "sub-system" of the speech generating equation (4).

Active Speech Recognition - Definition -

The active speech recognition is defined as "there exists active mode of speech recognition, if there exists the sub-system of the speech generating syatem in the speech receiving system when speech information is transfered completely."

Other Problems Treated by This Approach

Other problems which can be investigated by this general approach are the cocktail party problem, speech information transmission between normal and auditory-handicapped persons and others. A linear speech system will be shown on a slide at the meeting in order to make the story concrete and visible.

References

1) S. Inomata; JASJ meeting record (May 1960)

2) K. N. Stevens; JASA 32 1 (Jan. 1960)

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-14

PHONEMES ET PHONETOGRAPHE

Première partie : Loi de WEBER-FECHNER, compressions "idéales" et "réelles", phons et sones, tons et mels.

J.A. DREYFUS-GRAF
Ingénieur E.P.F.Z., Centre National d'Etudes des Télécommunications (C.N.E.T.), LANNION (France)

1 - L'existence physique des phonèmes

Les atomes sont nés dans l'imagination des philosophes antiques bien avant que les physiciens modernes aient pu les observer. Quant aux phonèmes -voyelles et consonnes- inventés par les linguistes, leur existence physique ne sera admise généralement qu'à partir du moment où ils seront reconnus complètement par des appareils automatiques, tels que le phonétographe [1], qui simulent certaines fonctions de l'oreille et du cerveau. Une difficulté provient du fait que diverses lois physiologiques fondamentales de la perception, sont encore controversées elles-mêmes, par exemple, celle de WEBER-FECHNER. [2]

2 - Les deux lois logarithmiques

Selon la loi de WEBER-FECHNER, l'intensité perçue x d'un signal varie comme le logarithme de l'amplitude S du signal émis, c'est-à-dire $x = \log S$. On peut se demander si, en fait, deux lois distinctes n'ont pas été confondues : D'abord, une loi simplement logarithmique ($x = \log_a S$), qui est l'inverse de la fonction exponentielle ($S = a^x$, avec une base $a>1$, telle que 2). Ensuite, une loi doublement logarithmique ($\log y = R.\log S$), qui est l'inverse de la fonction de puissance ($y = S^R$, avec un exposant $R<1$, tel que 0,1).

Dans la fonction exponentielle ($S = 2^x$), fondamentale en théorie de l'information, la base 2 est constante et l'exposant x est variable. Par contre dans la fonction de puissance ($y = S^R$), qui régit l'auto-régulation et la cybernétique, c'est l'exposant R qui est constant et la base S qui est variable. Nous allons examiner ces deux lois à l'aide de dispositifs techniques, et nous les comparerons aux courbes physiologiques des "sones" et des "mels". [3]

[1] DREYFUS-GRAF, J.A. - Phonétographe et cybernétique, 5ème C.I.A., A 36, LIEGE 1965
- Phonétographe IV : Progrès 1965 et programme de terminaison, Revue de Phonétique Appliquée, Mons, n° 1, 1966

[2] TRENDELENBURG, F. - Akustik, Springer-Verlag, BERLIN 1961

[3] BERANEK, L.L. - Acoustics, Mc Graw-Hill Book Co, NEW YORK 1954
ZWICKER E. & FELDTKELLER R. - Das Ohr als Nachrichten-Empfänger, Hirzel-Verlag 1967.

3 - Compresseurs "idéal" et "réel"

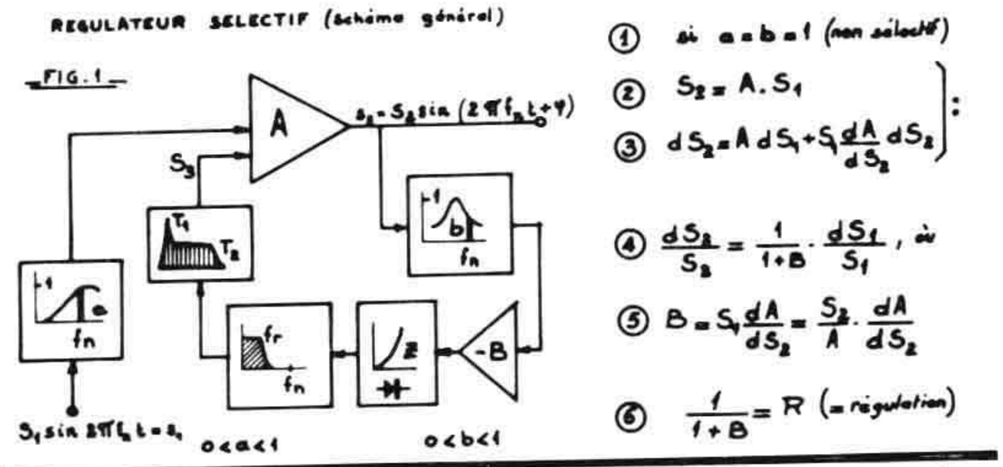
La figure 1 schématise un régulateur sélectif. Le signal d'entrée S_1 possède une fréquence f_n et une amplitude S_1, atténuée par "a" dans le filtre d'entrée. L'amplitude de sortie S_2 est égale à l'amplitude d'entrée S_1, multipliée par le "gain direct" A. Celui-ci est contrôlé par S_2 et par les paramètres suivants : l'atténuateur sélectif b, dû à un filtre de réaction, le gain de réaction-B, l'exposant de redressement z, la fréquence passe-bas f_r et les constantes de temps T_1, T_2, d'établissement et de retour du signal d'erreur S_3. En faisant abstraction de a=b=z=1, ce régulateur obéit aux équations (2) à (6), dûes à K. KUPFMULLER [4] Le compresseur d'amplitude "idéal" en est un cas particulier, avec compression R = constante, et en régime stationnaire.

Son extension "sélective" est illustrée par la Figure 2 et les équations (7) à (10). Toutefois, les compresseurs "réels", tels que ceux d'un phonétographe comprenant des tubes ou des semi-conducteurs, fournissent généralement des compressions R non-constantes, qui peuvent varier exponentiellement selon la Fig.2a et les équations (8a) à (10a).

Pour plus de clarté, les Figures 2 et 2 a peuvent être remplacées par nos "schémas logarithmiques ou cybernétiques" des Figures 3 et 3 a, avec les équations (11) à (19) et (11 a) à (19 a).[5]

On s'aperçoit alors que le compresseur "idéal" tend vers la loi doublement logarithmique (16) ou (18) et la fonction de puissance $S_2 = S_1^R$, tandis que le compresseur "réel" tend vers la loi simplement logarithmique (16 a) ou (18a) et la fonction exponentielle $S_1 = 2^{B.S_2}$. Dans ces équations, S_1= amplitude d'entrée, S_2 = amplitude de sortie, R = compression constante, B = "gain de réaction" $\simeq 1/R$.

Selon les équations (16 et (16a), l'atténuateur sélectif b affaiblit la compression tandis que l'exposant z de la courbe de redressement la renforce. Selon que z = 1, 2 ou 3, ce redressement est linéaire, quadratique, ou cubique, par exemple. Les paramètres a, b, B et z permettent de modifier profondément le fonctionnement d'un compresseur "réel", et nous allons voir dans quelle mesure on peut approcher ainsi celui d'un compresseur "idéal".



① si a = b = 1 (non sélectif)

② $S_2 = A.S_1$

③ $dS_2 = A\,dS_1 + S_1 \frac{dA}{dS_2} dS_2$:

④ $\frac{dS_2}{S_2} = \frac{1}{1+B} \cdot \frac{dS_1}{S_1}$, où

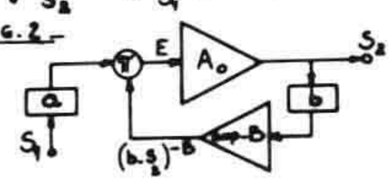
⑤ $B = S_1 \frac{dA}{dS_2} = \frac{S_2}{A} \cdot \frac{dA}{dS_2}$

⑥ $\frac{1}{1+B} = R$ (= régulation)

COMPRESSEUR SELECTIF "IDEAL"
COMPRESSION R = Constante $= \frac{1}{1+B}$

⑦ $\int \frac{dS_2}{S_2} = R \int \frac{dS_1}{S_1}$; $\log S_2 = R \log S_1 + C$

FIG. 2



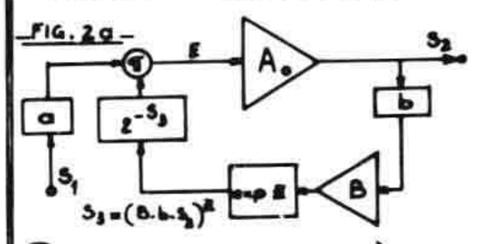
⑧ $S_2 = E.A_o$

⑨ $E = a S_1 : (b S_2)^B$ } x

⑩ $S_2 = (A_o a . S_1)^{\frac{1}{B+1}} : b^{\frac{B}{B+1}}$

COMPRESSEUR SELECTIF "REEL"
COMPRESSION R = variable $= 2^{-BS_2}$

FIG. 2a



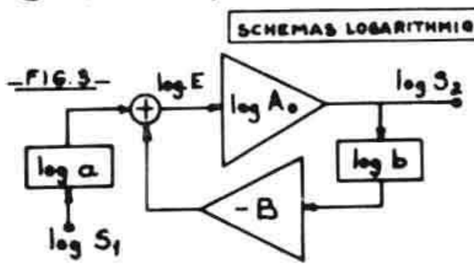
⑧a $S_2 = E.A_o$

⑨a $E = a S_1 . 2^{-(BbS_2)^z}$ } x

⑩a $S_2 = A_o a S_1 . 2^{-(B.b.S_2)^z}$

SCHEMAS LOGARITHMIQUES (CYBERNETIQUES) base 2

FIG. 3



⑪ $\log S_2 = \log E + \log A_o$

⑫ $\log E = \log a S_1 - B . \log b S_2$ } +

⑬ $\log S_2 = \frac{1}{1+B}(\log a S_1 + \log A_o) - \frac{B}{1+B} \log b$

⑭ $R = \frac{\Delta \log S_2}{\Delta \log S_1}$

⑮ si $a = A_o = 1$; $B \gg 1$; $bS_2 > 1$; $R = \frac{1}{1+B} \simeq \frac{1}{B}$

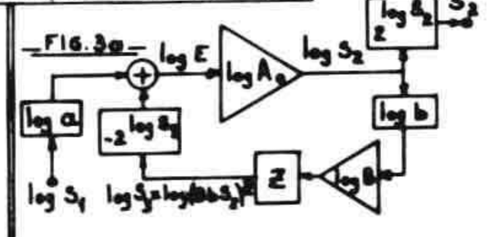
⑯ $\log S_2 \simeq \frac{1}{B} . \log S_1 - \log b$

⑰ si b = 1 (non sélectif) et $\frac{1}{B} \simeq R$

⑱ $\log S_2 \simeq R . \log S_1$ ⑲ $S_2 \simeq S_1^R = S_1^{1/B}$

fonction de puissance R

FIG. 3a



⑪a $\log S_2 = \log E + \log A_o$

⑫a $\log E = \log a S_1 - (bBS_2)^{BZ}$ } +

⑬a $(BbS_2)^z + \log_2 S_2 = \log_2 a S_1 + \log A_o$

⑭a (si $C[dB] \rightarrow R \simeq \frac{\Delta \log S_2}{\Delta \log S_1} = 1 + \frac{B.C}{8,7}$)

⑮a si $a = A_o = 1$ et $B.b \gg 1$:

⑯a $S_2 = \frac{1}{Bb}(\log S_1)^{1/z}$ ⑰a $S_1 \simeq 2^{(BbS_2)^z}$

⑰a si z = b = 1 (non sélectif) et $\frac{1}{B} \simeq r$:

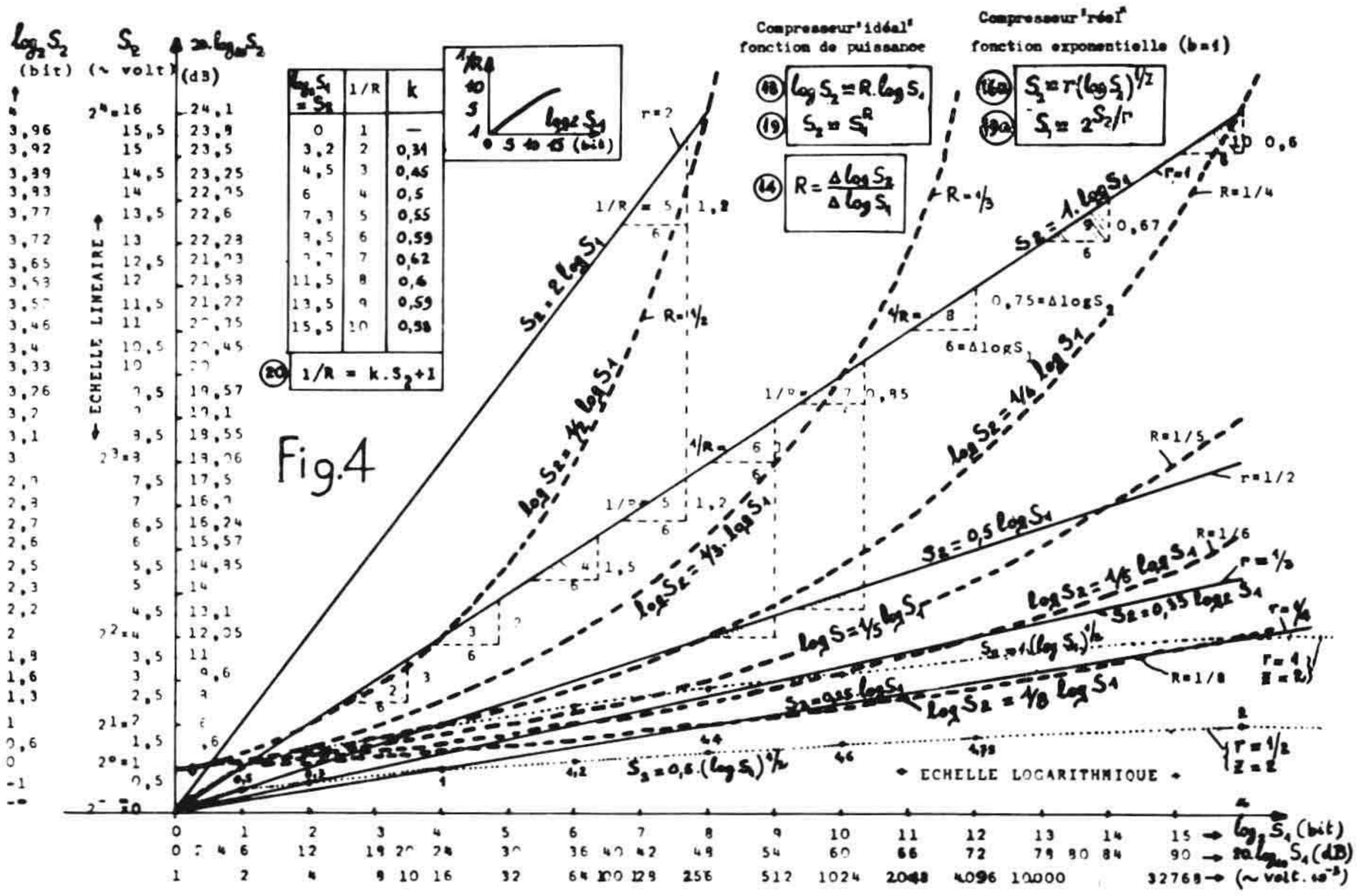
⑱a $S_2 \simeq r . \log_2 S_1$ ⑲a $S_1 \simeq 2^{S_2/r} \simeq 2^{BS_2}$

fonction exponentielle de S_2

B 24.4.68

[4] KÜPFMÜLLER, K. - Die Systemtheorie der elektrischen Nachrichtenübertragung, Hirzel-Verlag, Zürich-Stuttgart, 1952

[5] DREYFUS-GRAF, J.A. - Compresseurs d'amplitude et cybernétique, Bull. Techn.PTT, BERNE n° 7, 1964
- Spectres phonétographiques de la parole chantée, Festival International du Son, Editions Chiron, PARIS, mars 1968.
- Intensité sonore et auto-régulation, Revue du Son, n° 179, p.134-136; Editions Chiron, PARIS, mars 1968.



4 - Assimilation des deux lois divergentes

La figure 4 comporte une échelle horizontale logarithmique pour le niveau $\log S_1$ de l'amplitude d'entrée, et une échelle verticale linéaire pour l'amplitude de sortie S_2. Par conséquent, la loi doublement logarithmique du compresseur "idéal" (selon l'équation (18) $\log.S_2 = R.\log S_1$) se traduira par les six courbes en tirets, ayant $S_2 = 1$ pour origine, avec les six compressions $R = \frac{1}{2}, 1/3, 1/4, 1/5, 1/6, 1/8$. Par contre, la loi simplement logarithmique du compresseur "réel", selon l'équation (16a), $S_2 = r.\log S^{1/z}$ se traduira par cinq droites, en traits pleins, ayant $S_2 = 0$ pour origine, avec les cinq pseudo-compressions $r = 2, 1, 1/2, 1/3, 1/4$, étant sous-entendu que $z = 1$ (redressement linéaire). On constate ce qui suit : le compresseur "idéal" fournit une compression constante. Par exemple, avec $R = 1/4$, une variation du niveau d'entrée de 12 dB est toujours réduite à 3 dB à la sortie, quelle que soit la valeur du niveau d'entrée $\log S_1$. Par contre, le compresseur "réel" fournit, selon l'équation (20), une compression inverse $1/R = k.S_2 + 1$, qui varie avec S_2. Par exemple, $1/R$ varie entre 1 et 10 quand $\log S_1$ passe de 0 à 15,5 bits (ou de 0 à 93 dB), quelle que soit la valeur du paramètre r. Mais on s'aperçoit pourtant qu'il existe certaines valeurs et régions privilégiées pour lesquelles les deux lois divergentes se rapprochent remarquablement. Par exemple, la droite "réelle" S_2, avec $r = 1/3$ ($z=1$), peut être assimilée à la courbe "idéale" $\log S_2$, avec $R = 1/6$, tant que le niveau d'entrée $\log S_1$ reste compris entre 24 et 84 dB.

5 - Redressement non-linéaire :

Les deux courbes pointillées montrent l'effet d'un redressement quadratique ($z=2$) sur les droites en traits pleins, avec $r=1$ et $r=0,5$.

6 - Effet sélectif

L'équation (16 a) $S_2 = \frac{r}{b} (\log S_1)^{1/z}$ montre qu'en diminuant b tout se passe comme si on augmentait r, c'est-à-dire qu'on est libre de diminuer la compression en fonction de la fréquence f_n du signal s_1.

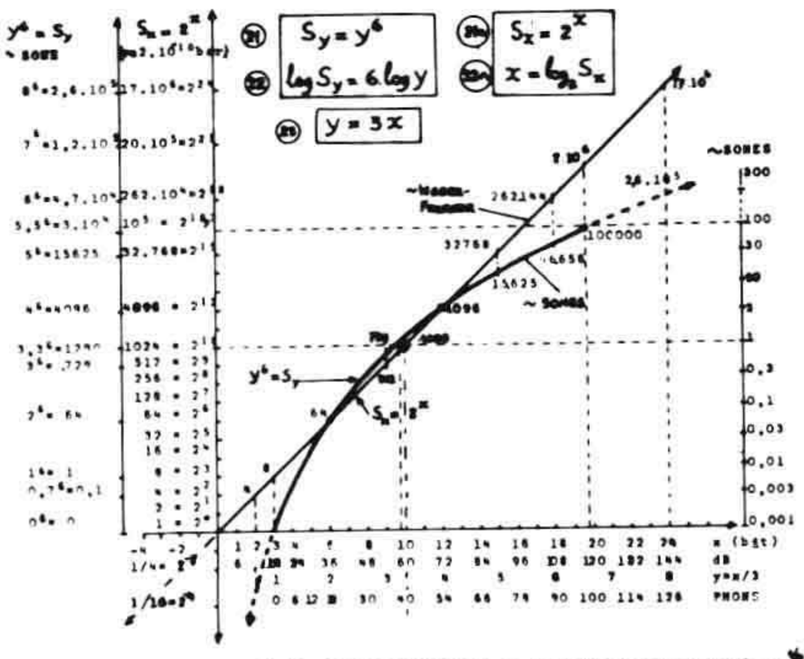


Fig.5. COMPARAISON ENTRE LA FONCTION DE PUISSANCE $S_y = y^{1/R}$ LA FONCTION EXPONENTIELLE $S_x = a^x$ ET L'ECHELLE DES "SONES", AVEC EXPOSANT 6 = 1/R (R=facteur de compression)
a = 2, x = 3y ; a=x^6.b

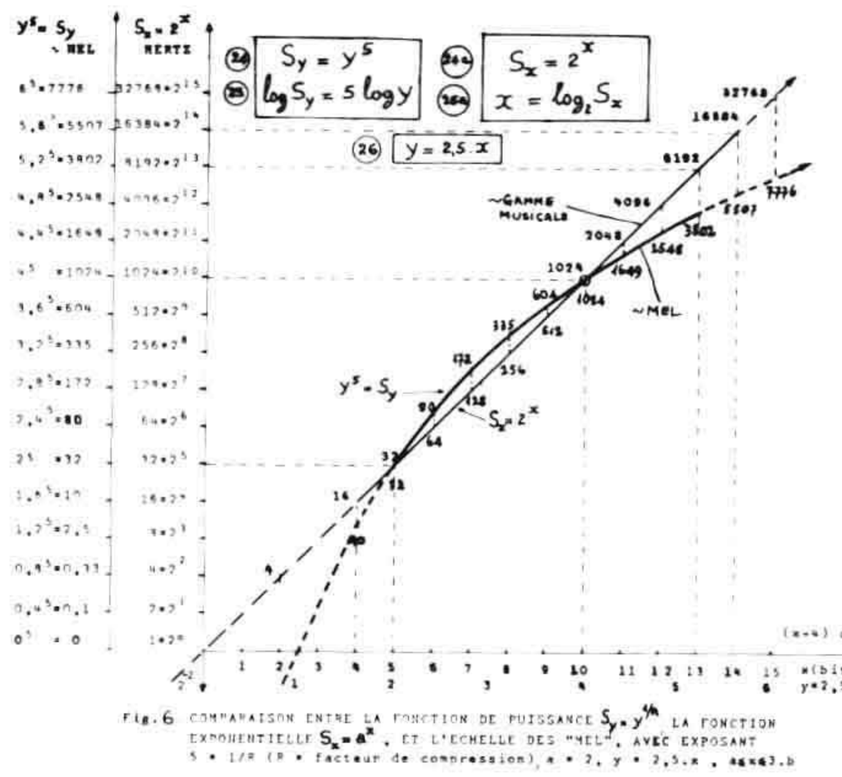


Fig.6 COMPARAISON ENTRE LA FONCTION DE PUISSANCE $S_y = y^{1/R}$ LA FONCTION EXPONENTIELLE $S_x = a^x$, ET L'ECHELLE DES "MEL", AVEC EXPOSANT 5 = 1/R (R = facteur de compression), a = 2, y = 2,5.x , a=x^3.b

7 - Compression "idéale" R = 1/6 et échelle des "sones"

Dans la Figure 5, la droite diagonale représente les niveaux de sortie $x = \log_2 S_x$ (échelle horizontale linéaire) en fonction des amplitudes d'entrée S_x (échelle verticale logarithmique), entre 0 et 24 bits, ou 0 et 144 dB. Cette droite, caractérisant à la fois un compresseur "réel" (avec a=b=z=r=1) et un décibel-mètre, coïncide avec la loi de WEBER-FECHNER. Opérons maintenant sur le diagramme les opérations suivantes : d'abord, nous faisons passer par les points 6 bits ($2^6 = 2^6$) et 12 bits ($2^{12} = 4^6$) la courbe $S_y = y^6$ (ou $\log S_y = 6.\log y$) d'un compresseur "idéal". Ensuite, nous translatons de 3 bits l'origine de l'échelle des bits, c'est à-dire de 18 dB celle des décibels. On s'aperçoit alors que l'échelle horizontale peut être graduée de 0 à 126 phons, et que la courbe mathématique du compresseur "idéal" est très proche de la courbe physiologique des "sones". Tout se passe comme si la nature avait réalisé dans l'oreille une compression constante, "idéale", $R \simeq 1/6$, des intensités sonores. Ainsi la loi de WEBER-FECHNER n'est valable physiologiquement qu'entre 18 et 54 phons. Entre 0 et 18 phons, il faudrait appliquer une loi dilatée $x \simeq 2.\log S$, puis, entre 54 et 126 phons, une loi contractée $x \simeq 0{,}5.\log S$.

8 - Compression "idéale" R=1/5 et échelle des "mels"

En expérimentant d'une manière analogue sur le diagramme de la Fig. 6, on découvre ce qui suit : la gamme musicale tempérée (des physiciens), coïncide avec la droite d'un compresseur "réel" $x = \log_2 S_x$ (ou $S_x = 2^x$). Par contre, la courbe mathématique d'un compresseur "idéal" $S_y = y^5$ (ou $\log S_y = 5.\log y$), passant par les tons musicaux 32 Hertz ($2^5 = 2^5$) et 1024 Hertz ($2^{10} = 4^5$), se rapproche remarquablement de la courbe physiologique des "mels". Tout se passe comme si la nature avait réalisé dans l'oreille une compression constante, "idéale", $R \simeq 1/5$, des hauteurs mélodiques. Ainsi, la gamme tempérée n'est valable physiologiquement qu'entre 80 et 800 Hertz environ. Entre 16 et 80 Hertz, il faudrait appliquer une gamme dilatée $x \simeq 1{,}6.\log_2 S$, puis entre 800 et 16000 Hertz une gamme contractée $x \simeq 0{,}6.\log_2 S$.

Remarque : Les compressions 1/5 et 1/6 ci-dessus ne sont valables qu'à titre indicatif, de même d'ailleurs que les courbes des "sones" et des "mels", qui sont encore à clarifier.[3]

Il appartiendra à des études ultérieures de préciser les valeurs des divers paramètres, d'examiner l'effet des transitoires et des sons complexes, ainsi que les autres aspects des sujets "phonèmes et phonétographe".

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-15

CONSONANTAL CLUSTERS AND CONNECTED SPEECH RECOGNITION

D. R. Reddy

Assistant Professor of Computer Science

Stanford University

Stanford, California

U.S.A. 94305

INTRODUCTION

Consonantal clusters are usually considered to be more difficult to analyze than consonant-vowel sequences. One possible source of difficulty arises from the competence and performance of the speaker. Because consonantal clusters are harder to pronounce a speaker may omit or mispronounce some of the phonemes making it difficult to determine what was intended. The other possible difficulty is the ever present segmentation problem. Given that a phoneme sequence is a consonantal cluster the question arises if it is indeed harder to determine phoneme boundaries for this sequence than for any other sequence. Our results indicate that, if the speaker is competent and cooperative, the analysis of consonantal clusters is no harder than any other sequence. This paper presents the results of computer analyses of some connected speech segments to illustrate the problems involved.

MATHOD AND MATERIALS

About 40 utterances with initial, medial and final consonantal clusters were analyzed using a time-shared PDP-6 computer with 65K words of memory. Audio input consisted of a microphone connected to the computer via an A-D converter. The speech signal was sampled at 20000 samples per second and digitized to 9-bit samples. The digitized waveform was segmented using an extension of the procedure described by Reddy (1966). Smoothing and differencing operations on the waveform generate parameters which are used to determine whether the characteristics of the sound are changing or similar, generating transitional and sustained segments respectively. A classification procedure labels the segments as vowel, fricative, stop, etc. The waveform traces given in the following section illustrate the results of the segmentation procedure and were graphed using a CALCOMP plotter.

RESULTS AND DISCUSSION

Whether the features of two adjacent phonemes are distinguishable or not will depend on the similarity of the two phonemes and the possibility of coarticulation. Table 1 (Whitney, 1896) formulated by Sanskrit grammarians over two thousand years ago, provides a simple way of determining whether two phonemes are close. The entries are so arranged that the closer the phoneme the smaller is the distance. In particular, elements in the same row (or same column) are more likely to be confused with each other than elements in different rows (or different columns). Table 1 contains most of the sounds of English, and can be used to determine phonemic similarities with little modification.

		Gutt.	Pal.	Ling.	Dent.	Lab.	
Son.	Vowels	a, ā (19·78 8·19)					
			āi, e (·51 2·94)		o, āu (1·88 16)		
			i, ī (4·85 1·19)	ṛ, ṝ (·74 ·01)	ḷ u, ū (·01 2·61 ·73)		
	Semivowels		y (4·25)	r (5·05)	l (·09)	v (4·99)	
	Nasals		n̄ (·22)	n (·35)	ṇ (1·03)	n (4·81)	m (4·34)
	Anusvāra	ṁ (·63)					
	Aspiration	h (1·07)					
Surd	Visarga	ḥ (1·31)					
	Sibilants		ç (1·57)	ṣ (1·45)	s (3·65)		
Son.	Mutes, asp.		gh (·15)	jh (·01)	ḍh (·03)	dh (·83)	bh (1·27)
	Mutes, unasp.		g (·82)	j (·94)	ḍ (·21)	d (2·85)	b (·45)
Surd	Mutes, asp.		kh (·13)	ch (·17)	ṭh (·06)	th (·58)	ph (·03)
	Mutes, unasp.		k (1·99)	c (1·26)	ṭ (·26)	t (6·65)	p (2·46)

Table 1. Grouping of Sanskrit Alphabet

Figures 1 and 2 illustrate the possible loss of a phoneme in consonantal clusters if two adjacent phonemes are the same or similar. Figure 1 shows a portion of the waveform of some milk. The ending /m/ of some and the beginning /m/ of milk are

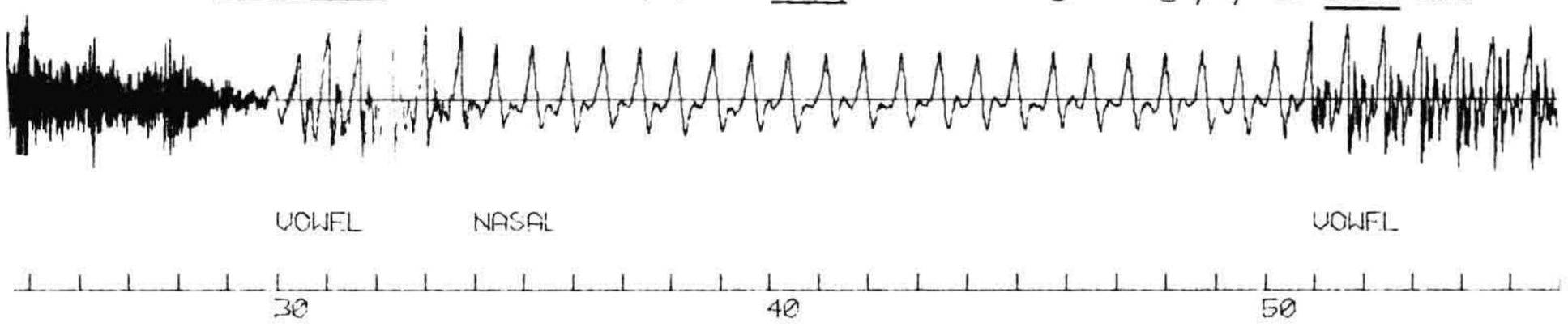


Figure 1. Waveform of part of the sound 'some milk'

grouped together as one segment. Perhaps duration of the segment could be used to indicate the possibility of an extra phoneme, but it is not a reliable feature. Another example of a similar type is the loss of /t/ in I hate to cook. Figure 2 illustrates the loss of a phoneme in the phrase solved it. /v/ and /d/ are similar

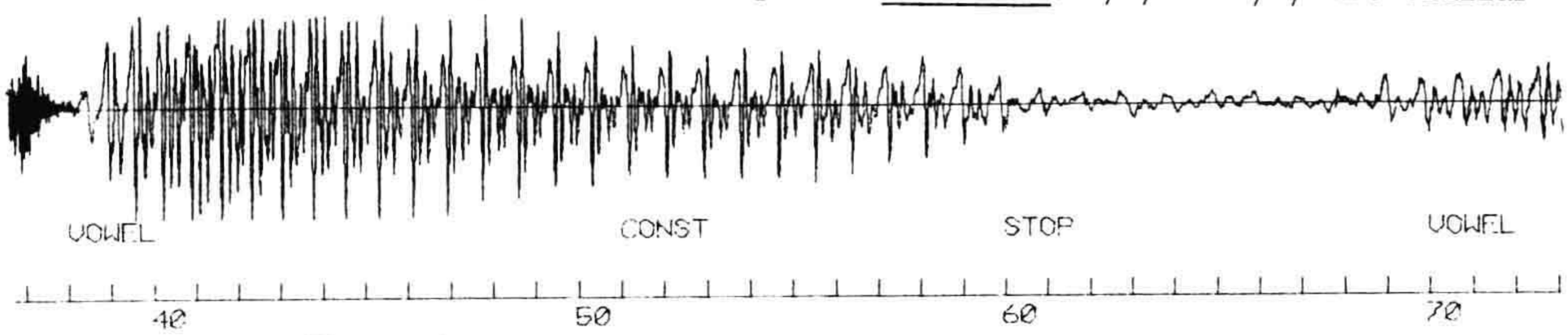


Figure 2. Waveform of part of the sound 'solved it'

and were grouped together. The formant transitions could be used to resolve the ambiguity, but then all the elements in the labial column of Table 1 would have

similar transitions. More detailed analyses would certainly help, but one of the critical problems of speech recognition is determining where detailed analyses are needed and where they are not.

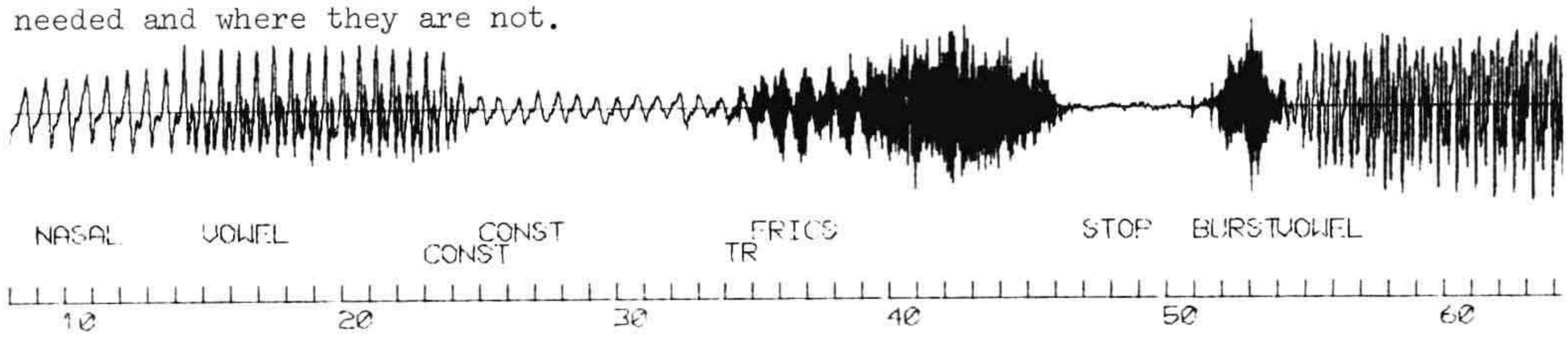


Figure 3. Waveform of part of 'midst of'

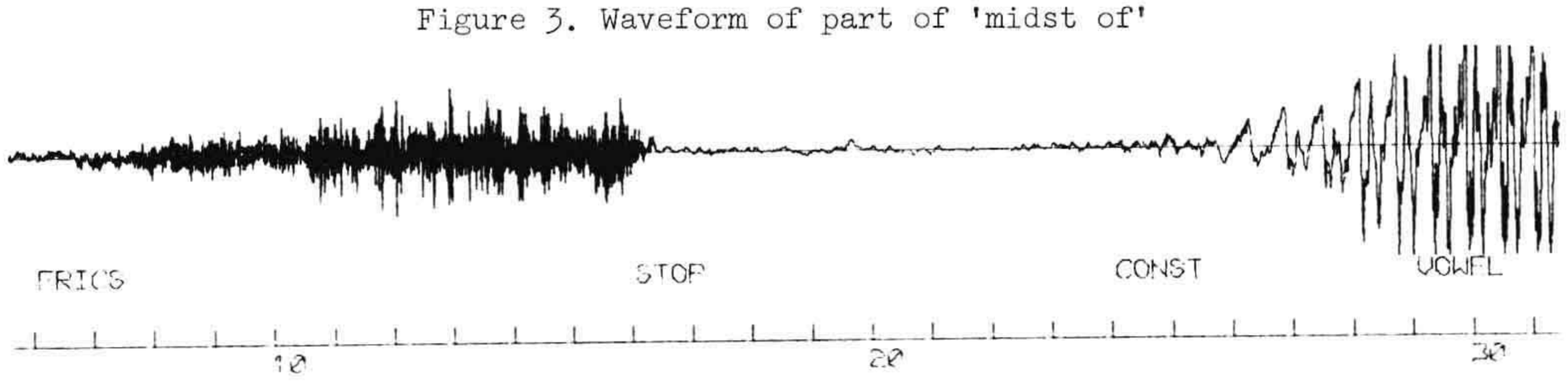


Figure 4. Waveform of part of 'split'

Not all consonantal clusters are as problematic as those in Figures 1 and 2. Figures 3 and 4 illustrate two instances where phoneme boundaries are well defined. Figure 3 shows the final CC of the word <u>midst</u> and Figure 4 shows the initial CC of the word <u>split</u>. Many CC belong to this category and do not pose any special problems.

The other major difficulty with consonantal clusters results from coarticulation. The articulators, viz., lips, tongue, soft palate, and the vocal cords, can all operate in parallel occasionally resulting in the overlap of features of adjacent phonemes. In Figure 5, the usual silence preceding the burst of /t/ in the word <u>seventeen</u> is almost missing because the tongue is already in position as a result of

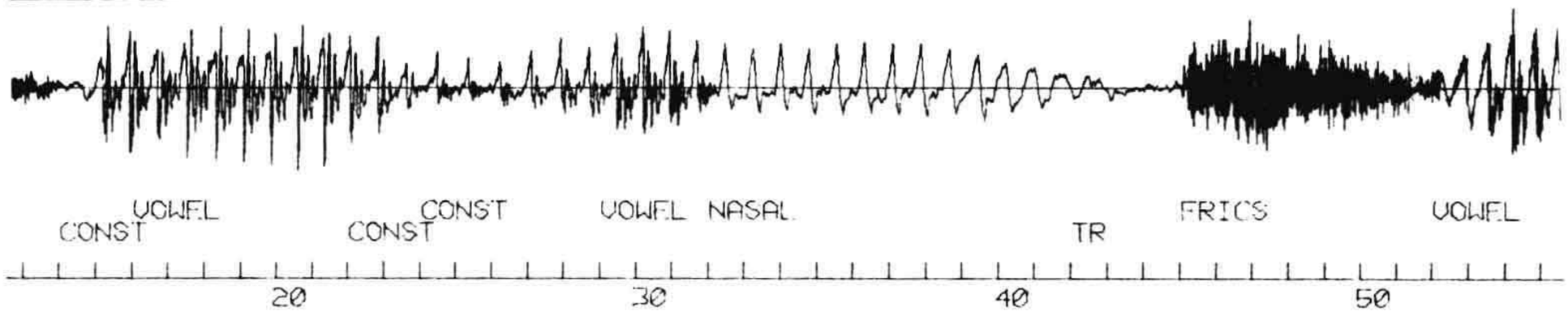


Figure 5. Waveform of part of 'seventeen'

the preceding /n/. In Figure 6, the /h/ in <u>John has</u> merely modifies the characteristics of the preceding /n/ and the following /æ/, and has not existence of its own.

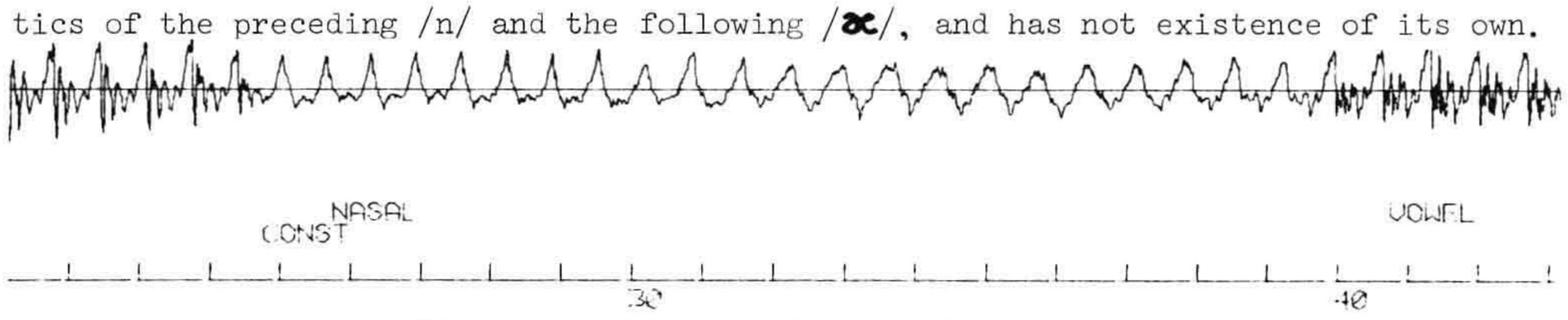


Figure 6. Waveform of part of 'John has'

The last example is much more serious than the preceding one because, instead of a mere loss of some of the features of one phoneme, the features of both phonemes are intermixed generating a new set of features. This problem is analogous to the recognition of two people speaking at once. There exist no simple solutions.

The problems outlined above are by no means unique to consonantal clusters. In the phrase she is nice the /i/ at the end of she is often combined with /I/ at the beginning of is resulting in the loss of a phoneme. In some respects CC are better than VCV sequences for speech recognition. In Figure 7, the /t/ in what is

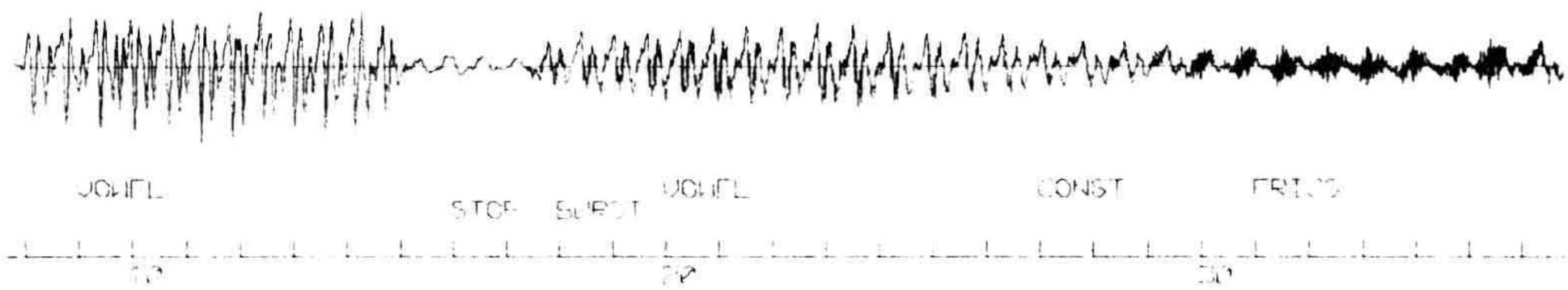


Figure 7. Waveform of part of 'what is'

has neither a noticeable silence, nor a burst, and is in fact voiced. This is because, in connected speech, the influence of neighbors on the features of a phoneme is greatest when the phonemes are very dissimilar. This raises a curious paradox: if two neighboring phonemes are similar then it is hard to segment and if they are very dissimilar they modify each others' parameters so much that they are hard to classify. Fortunately for us most speech occupies the in-between ground.

CONCLUSIONS

In spite of all the above examples, phonemic representation of speech is the best model we have available. Rather than discarding it we should add some exception handling features. Consonantal clusters pose some problems but no more than any other sequence of phonemes. It would be desirable to develop rules for transformation of phonemic features to indicate how they might be modified in a particular context of neighboring phonemes.

ACKNOWLEDGEMENTS

The research reported here was supported by the Advanced Research Projects Agency of the Department of Defense.

REFERENCES

1. Reddy, D. R., Segmentation of Speech Sounds, J. Acoust. Soc. Am., 40, 307-312 (1966).

2. Whitney, W. D., Sanskrit Grammer, Ginn and Co., Boston, Mass. (1896).

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-17

The Optimum Filter for Frequency Analysis of Speech

ISOMICHI- Yoshinori
Electrotechnical Laboratory

1. Introduction

When performing frequency analysis of speech sound, the time-resolution becomes worse with the improvement of frequency-resolution. This phenomenon is known as the uncertainty relation of filters. In this paper, some of new results of the optimum filter for speech analysis have been obtained.

2. Gabor's result-Gaussian filter [1] [2]

The above problem was investigated by D. Gabor. He borrowed the next relation from the uncertainty principle in quantum mechanics;

$$\frac{\int_{-\infty}^{\infty}(t-t_0)^2|g^+(t)|^2dt}{\int_{-\infty}^{\infty}|g^+(t)|^2dt}\cdot\frac{\int_{-\infty}^{\infty}(f-f_0)^2|G^+(2\pi if)|^2df}{\int_{-\infty}^{\infty}|G^+(2\pi if)|^2df}\geqq\left(\frac{1}{4\pi}\right)^2 \tag{1}$$

Where, $$G^+(2\pi if)=\int_{-\infty}^{\infty}g^+(t)e^{2\pi ift}dt \tag{2}$$

The equatliy holds only when

$$g^+(t)=\frac{A}{(2\pi(\Delta t)^2)^{\frac{1}{4}}}e^{-\frac{(t-t_0)^2}{4(\Delta t)^2}}e^{i(2\pi f_0t+\theta)} \tag{3}$$

$$G^+(2\pi if)=A(8\pi(\Delta t)^2)^{\frac{1}{4}}e^{-(2\pi)^2(\Delta t)^2(f-f_0)^2}e^{-2\pi i(f-f_0)t_0+i\theta} \tag{4}$$

Where, A is the energy of $g^+(t)$, Δt is the mean duration of $g^+(t)$ and θ is an arbitrary real number.

Using the above result, he obtained the impulse response of the optimum filter as

follows;

$$g(t)=\{g^{+}(t)+\overline{g^{+}(t)}\}/2 \tag{5}$$

$$=\frac{A}{(2\pi(\Delta t)^2)^{\frac{1}{4}}}\,e^{-\frac{(t-t_0)^2}{4(\Delta t)^2}}\cos(2\pi f_0 t+\theta) \tag{6}$$

Its transfer function can be given using the relation,

$$G(2\pi if)=\{G^{+}(2\pi if)+\overline{G^{+}(-2\pi if)}\}/2 \tag{7}$$

This type of filter is called the Gaussian filter because its impulse response has the same form as the Gaussian distribution in the statistics.

The Gaussian filter is optimal in Gabor's sense. However, it can not be regarded as a true optimal filter from the next two points. First, the impulse response spreads over the negative time axis. Therefore it can not be a physically realizable filter. Secondary, the relation between g(t) and $g^{+}(t)$ has not been explained satisfactorily. (The function $g^{+}(t)$ is not the analytic signal of g(t), because the Fourier transform of $g^{+}(t)$ does not vanish for negative frequency.

3. Analytic Spectrum Method [3]

We will suppose that g(t) is defined only in the interval $(0,\infty)$. The spectrum of g(t) is defined by the Laplace transform of g(t).

By the analytic spectrum method, the "positive spectrum" of G(S) can be written as,

$$G^{+}(S)=\frac{-1}{\pi i}\,\text{v.p.}\int_c \frac{G(z)}{z-S}\,dz \qquad \text{on} \begin{cases} Re\{S\}>0 \\ \text{or} \\ Im\{S\}<0 \end{cases} \tag{8}$$

Where, the contour c is $(-\infty \longrightarrow 0 \longrightarrow i\infty)$

And also, the next relation can be obtained;

$$G(S)=\{G^{+}(S)+\overline{G^{+}(\bar{S})}\}/2 \tag{9}$$

Letting g(t), $g^{+}(t)$ be the Laplace inverse transforms of G(S), $G^{+}(S)$ respectively, we can obtain

$$g^{+}(t)=\begin{cases} \frac{1}{\pi i}\,\text{v.p.}\int_c G(S)e^{st}\,ds & (t\geq 0) \\ 0 & (t<0) \end{cases} \tag{10}$$

$$g(t)=\{g^{+}(t)+\overline{g^{+}(t)}\}/2=Re\{g^{+}(t)\}. \tag{11}$$

Where, $G^{+}(S)$ is an analytic function which is regular on the three quarters of s-domain.

The positive spectrum $G^{+}(S)$ given above is different from $G^{+}(S)$ given by analytic

signal method. (The latter is zero on the lower half plane of s-domain and it is 2G(S) on the upper half plane.)

4. Definitions of the characteristics of a filter

The following characteristics are defined in the sa way as the "analytic signal" method.

1. Mean delay of filter, t_0

$$t_0 = \int_0^\infty t|g^+(t)|^2 dt \Big/ \int_0^\infty |g^+(t)|^2 dt \tag{12}$$

2. Mean duration of filter, Δt

$$(\Delta t)^2 = \int_0^\infty (t-t_0)^2 |g^+(t)|^2 dt \Big/ \int_0^\infty |g^+(t)|^2 dt \tag{13}$$

3. Absolute delay of filter, τ_0

$$\tau_0 = \underset{\tau}{SUP}\{\tau;\ g^+(t) \equiv 0,\ for\ t<\tau\} \tag{14}$$

4. Absolute effective duration of filter, T

$$T = \underset{\tau}{Inf}\{\tau;\ g^+(t) \equiv 0,\ for\ t>\tau\} - \tau_0 \tag{15}$$

5. Time constant of filter, $\Delta\tau$

$$(\Delta\tau)^2 = \int_{\tau_0}^\infty (t-\tau_0)^2 |g^+(t)|^2 dt \Big/ \int_{\tau_0}^\infty |g^+(t)|^2 dt \tag{16}$$

6. Mean frequency of filter, f_0

$$f_0 = \int_{-\infty}^\infty f|G^+(2\pi i f)|^2 df \Big/ \int_{-\infty}^\infty |G^+(2\pi i f)|^2 df \tag{17}$$

7. Mean band width of filter, Δf

$$(\Delta f)^2 = \int_{-\infty}^\infty (f-f_0)^2 |G^+(2\pi i f)|^2 df \Big/ \int_{-\infty}^\infty |G^+(2\pi i f)|^2 df \tag{18}$$

As a whole, the analytic spectrum method can be said to give more natural definitions than the analytic signal method does.

5. The optimum filter for frequency analysis

5;1. The case where absolute duration is not bounded

In this case we use $\Delta\tau\cdot\Delta f$ for the measure of uncertainty. Using the variational method, we can obtain

$$\Delta\tau\cdot\Delta f \geq \frac{3}{4\pi} \tag{19}$$

The equality holds, only when

$$g^+(t) = \begin{cases} 2A\left(\frac{3}{2\pi(\Delta\tau)^2}\right)^{\frac{1}{4}}\left(\frac{3}{2(\Delta\tau)^2}\right)^{\frac{1}{2}}(t-\tau_0)e^{-\frac{3}{4(\Delta\tau)^2}(t-\tau_0)^2}e^{i(2\pi f_0 t+\theta)} & (t \geq \tau_0) \\ 0 & (t<\tau_0) \end{cases} \tag{20}$$

Where, A is the energy of $g^+(t)$, $\Delta\tau$ is the time constant of $g^+(t)$ and θ is an arbitrary real number.

Therefore the impulse response of the optimum filter can be given as,

$$g(t)=\begin{cases} 2A\left(\frac{3}{2\pi(\Delta\tau)^2}\right)^{\frac{1}{4}}\left(\frac{3}{2(\Delta\tau)^2}\right)^{\frac{1}{2}}(t-\tau_0)e^{-\frac{3}{4(\Delta\tau)^2}(t-\tau_0)^2}\cos(2\pi f_0 t+\theta) & (t\geq\tau_0) \\ 0 & (t<\tau_0) \end{cases} \tag{21}$$

The transfer function exists but can not be expressed explicitly.

;2. <u>The case where absolute duration is bounded.</u>

In this case we use $T\cdot\Delta f$ for the measure of uncertainty. Using the variational method we can obtain

$$T\cdot\Delta f\geq\frac{1}{2}. \tag{22}$$

The equality holds only when

$$g^+(t)=\begin{cases} \sqrt{\frac{2}{T}}\,A\sin\left\{\frac{\pi(t-\tau_0)}{T}\right\}e^{i(2\pi f_0 t+\theta)} & (\tau_0\leq t\leq T+\tau_0) \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

Where, A is the energy of $g^+(t)$, τ_0 is the absolute delay of $g^+(t)$ and θ is an arbitrary real number.

The transfer function of $g^+(t)$ can be written as

$$G^+(2\pi if)=\frac{\sqrt{2T}\,A}{\pi}\,\frac{\left\{e^{-(T+\tau_0)2\pi i(f-f_0)}+e^{-\tau_0 2\pi i(f-f_0)}\right\}e^{i\theta}}{1-4(f-f_0)^2T^2} \tag{24}$$

Using the above relation, the impulse response of the optimum filter is given by

$$g(t)=\sqrt{\frac{2}{T}}\,A\sin\left\{\frac{\pi(t-\tau_0)}{T}\right\}\cos(2\pi f_0 t+\theta)\qquad(\tau_0\leq t\leq T+\tau_0) \tag{25}$$

And its tranfer function can be obtained by (7).

<u>References</u>

[1] Gabor, D.
Jour. Inst. Elec. Engrs.
Vol. 93, Pt. 3, P.P. 429 ~ 441 (1946)

[2] Kay, I. and Silverman, R.A.
Information and Control
Vol. 1, P.P. 64 ~ 75 (1957)

[3] Isomichi-Y and Iijima, T.
Trans. Inst. Elec. Conmum. Engrs. (Japan)
Vol. 51 - C, No. 3, P.P. 83 ~ 90 (1968)

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-18

Real-Time Colored Display for Sound Spectrograms

Tsuneji Koshikawa, Yasuaki Nakano,
Takeshi Nakayama, and Tanetoshi Miura
Central Research Laboratory, Hitachi, Ltd.

Introduction

We built on trial a real-time colored display for sound spectrograms. This apparatus was aimed at being an effective means for recognition of various acoustic signals.[1)]

One of the main features of this apparatus is that the spectrogram is displayed on a color picture tube in real-time in a colored contour form resembling a contour map by converting the spectrum level to color (11 classes).

Another feature is that the spectrum pattern can be seen in sky-sigh fashion by shifting the instantaneous spectrum on time axis reserving it in a certain time interval.

Construction of apparatus

The apparatus consists of two main parts. One is a real time spectrograph and the other a colored display which displays the output of the spectrograph on a color picture tube.

For the spectrograph, we adopted Gill's system[2)] which uses time scale compression and heterodyne frequency analysis. With this analyzer, the time series of spectrum section are obtained as output signal.

The number of analyzing channels is 100, and equivalent analyzing band width is set at 125 Hz or so. Sweeping time for full frequency range (100～6000 Hz) is 8 msec.

(time length per spectrum section).

The colored display part consists of (1) coder and decoder which convert the eleven quantized spectrum levels to color, (2) time scale compressor for sky-sign-like display and (3) color picture tube and surrounding circuits. (Fig. 1)

We used a logarithmically equal interval (4 dB), as the quantizing interval of spectrum level, in order to cover the dynamic range of about 40 dB. Of course, the quantizing interval can be decided optionally according to the purpose of display.

The final form of spectrogram displayed on the color picture tube is presented as follows. The ordinate designates frequency (100 steps) and the abscissa designated time. The spectrum amplitude is presented by hue (11 classes), and every spectrum section runs for the duration of 0.8 sec. from right to left.

Furthermore, it is capable of displaying the spectrogram using not only hue but also brightness and hue or brightness alone (i.e. monochromatic display).

Method of color arrangement

The arrangement of hues for the eleven steps of spectrum level was determined as follows referring to the visual effect and the familiar map system. Namely, red, an advancing color, was assigned to the highest level, blue, a retiring color, was assigned to the lowest level, and nine hues were selected between red and blue in sensorily equal intervals. The chromaticity value of each hue on the x-y chromaticity diagram is shown in Fig. 2.

Principle of display of sky-sign-like fashion

First, each spectrum section (upper part of Fig. 3) having the duration of T is delayed repeatedly (delay time is $T(N-1)/N$), then the time length of each spectrum section is compressed to 1/N. Next, time compressed sections 1', 2', ... are cut out in intervals of T. (middle part of Fig. 3). Finally, these spectrum sections are delayed again and replaced in the form as shown in the lower part of Fig. 3. Then spectrum sections 1', 2', ... are shifted successively in time slots (1), (2), ... Therefore, moving spectrograms are displayed, the horizontal axis sychronized with T and vertical axis with T/N.

In this apparatus, T=8 msec. and N=100; thus time length for sky-sign-like

display is 0.8 sec.

Examples of display and some applications

A few typical colored spectrograms are shown in slide and cine-film. It is confirmed that colored display of this kind is more useful than usual brightness-modulating display for the visual discrimination of signal differences.

For this kind of display, various applications may be expected, e.g., signal detection and recognition in Sonar systems, observation of frequency characteristics of reverberation in concert halls, music education and so on.

Concluding remarks

Presented above is the prototype of a new display apparatus in which color is used effectively. We would like to pursue further optimum and convenient systems for various applications and developments.

Acknowledgment

The authors wish to acknowledge the helpful suggestions and instructions offered by Mr. K. Nirei, director, Dr. Y. Ito, and Dr. I. Shinbo, all of the 5th Research Center, Technical Research and Development Institute, Japan Defence Agency; and to express their thanks to Dr. T. Anzai, senior engineer at Totsuka Works, Hitachi, Ltd., for his valuable advice and support.

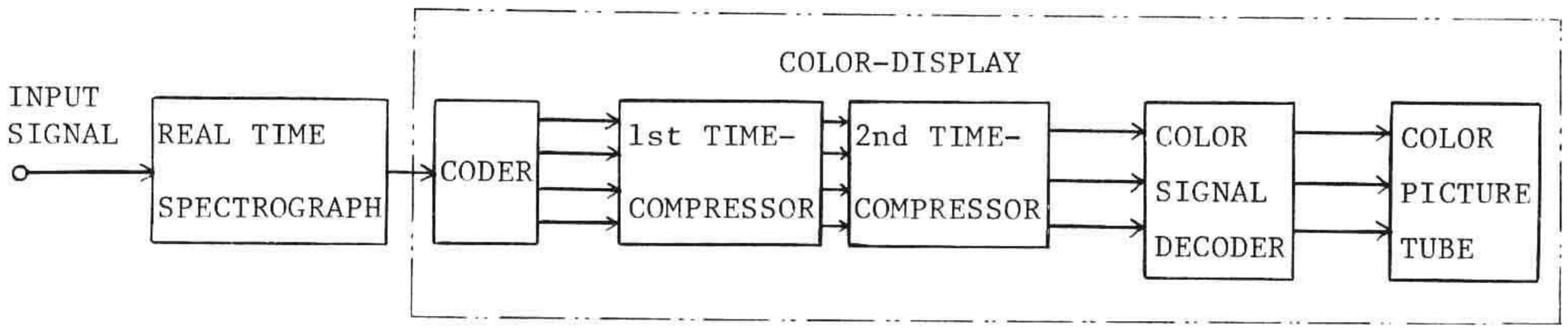


Fig. 1 Block diagram of real-time colored display of sound spectrogram

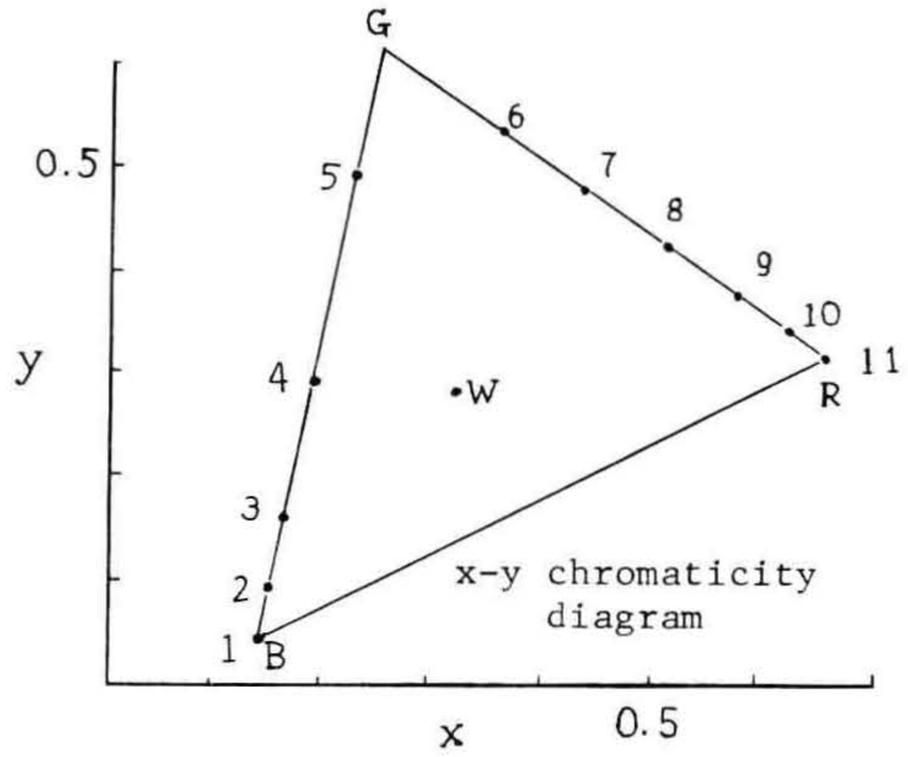


Fig. 2 Eleven points of chromaticity selected for colored contour display

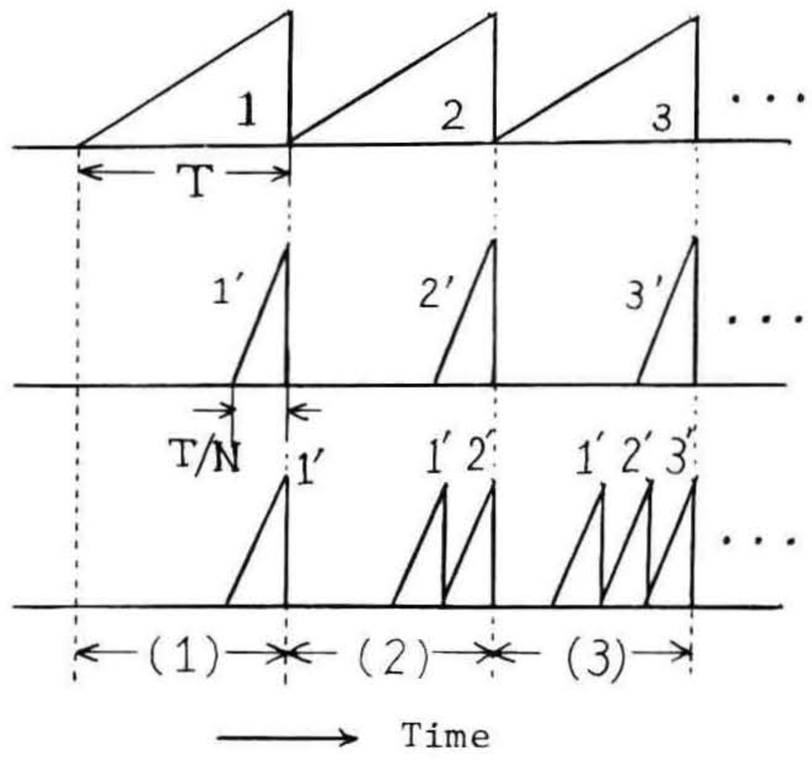


Fig. 3 Principle of sky-sign-like display

References

1) Yasuaki Nakano et al.; 74th meeting of A.S.A., Nov. 1967

2) J.S. Gill; Nature 4759, 117, Jan. 1961

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

C-5-19

Analysis of Speech Sounds in Ultrasonic Frequency Range

L. L. Miasnikov
Professor of Leningrad Shipbuilding Institute
H. H. Miasnikova
Docent of Leningrad Mechanical Institute
N. P. Gromov
Engineer of Leningrad Shipbuilding Institute

Introduction

The consonants are generated by the air stream passing through some openings in the vocal tract; the sound sources which appear in the process of articulation of consonants in the majority of cases are of air-jet or whistle type. As it was found in our investigations, the speech organ plays additionally the part of an ultrasonic whistle, especially in the case of fricatives s, $\int$; it produces high frequency, ultrasonic speech components of measurable sound pressure levels. In ultrasonic high frequency range some resonances appear. The principle resonance maximum close to $80^{kc/s}$ is noticeable for s and z and is less marked for f, h, k, when the ultrasonic resonator formed between lips and tongue is deranged. The ultrasonic emission occurs also by articulation of labiodental , compound consonants ts and $t\int$ etc. It is obvious that for all vowels the ultrasonic spectral components, if they exist, are negligible.

Spectral Analysis of Speech Sounds

The experimental set consisted of titanat barium receiver with a calibrated amplifier, covering the frequency range includ-

ing the frequency interval from 20 till 260 kc/s and a Freistedt's spectrograpf with 1/3 octave band pass filters, the system of filters covering the frequency range already mentioned. The 0 db level corresponds to a sound pressure level of 0,05 n/m².

The spectral analysis of principal Russian and English phonemes was realized with the participation of 5 speakers which kept the distance 0,15 m from the receiver. They pronounced vowells and fricatives in a drawling manner but others consonents in connection with an **e** ending (e.g. "pe", "be" and so on). More than 1000 spectrogrammes were obtained.

Some results of ultrasonic spectral analyses of English consonants are represented in Fig.1, the 0 db level given above. The data prove the existance of ultrasonic speech components. Really there exists a maximum lying close 80 kc/s. The spectra of f and θ show a considerable amount of high frequency components. The investigation of sound spectra of f and θ (soft th) showed a markable difference in ultrasonic region. For f one

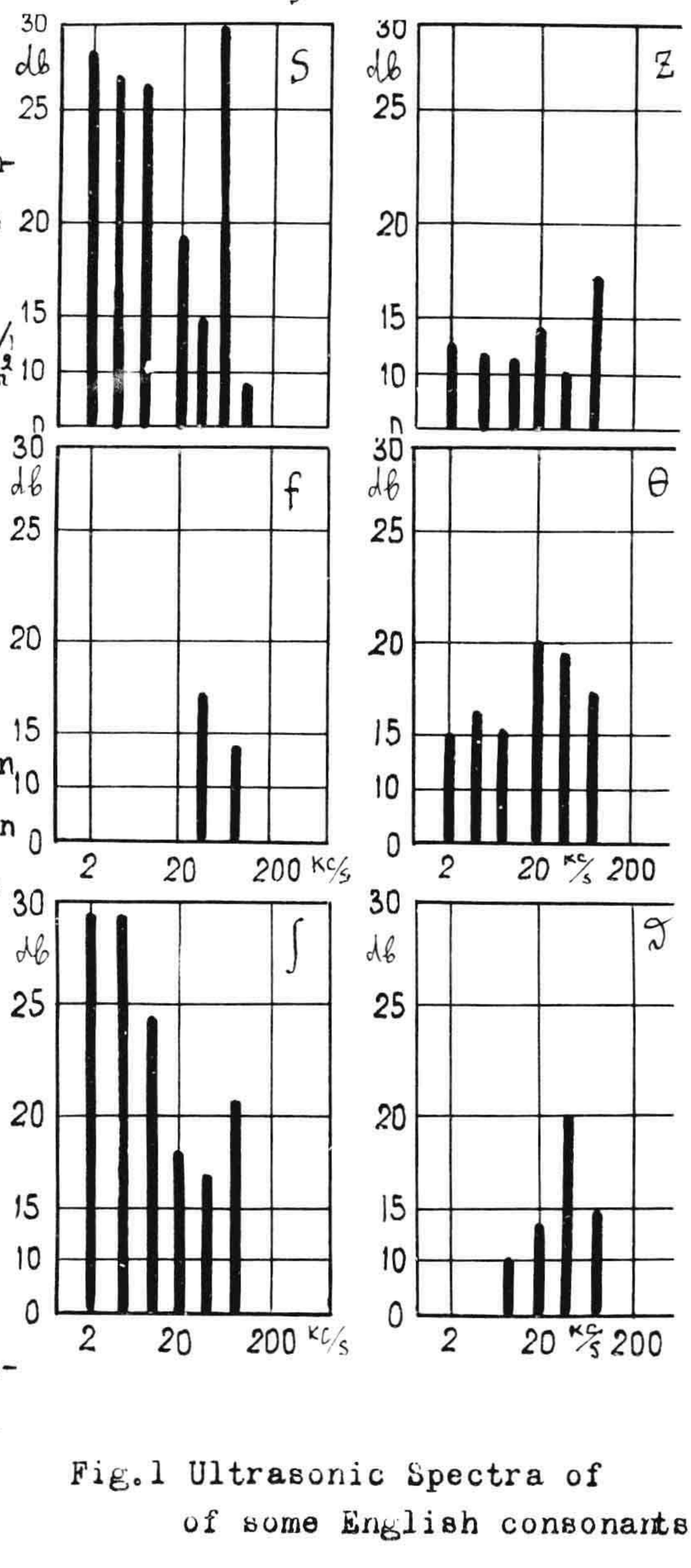


Fig.1 Ultrasonic Spectra of some English consonants

may observe a rather narrow spectral band near 80 kc/s; for θ this band is spread more widely. The difference which may be important for the problem of automatic recognition of f and θ occures, perhaps, due to an ultrasonic resonator between lips and teeth in the case of f articulation, dropping for θ , when the volume reduces and only a gap between upper teeth and the end of the tongue is acting. The results of spectral analysis of Russian phonemes are given in Table 1.

Table 1. Ultrasonic Speech Components

(0 db corresponds to 0,05 n/m^2)

Phonemes		Sound Pressure levels (db) for frequencies (kc/s):			
		20	40	80	160
Open consonants (fricatives)	f	-	17	14	-
	v	5	5	-	-
	s	22	18	27	1
	z	13	10	18	-
	ʃ	18	17	20	-
	ʒ	18	-	-	-
Palatial	h	5	-	16	-
	g	-	-	-	-
	k	5	-	10	-
Stops	p	5	-	-	-
	b	-	-	-	-
	t	15	10	-	-
	d	-	-	-	-
Compound	ts	22	16	13	-
	tʃ	10	10	5	-
Vowels		-	-	-	-

Analysis of Speech Sounds in Ultrasonic Frequency Range

The existance of ultrasonic components in the spectra of several phonemes is significant for automatic speech recognition problem [1-3]. As it is well known this problem meets considerable difficulties due to overlapping of phoneme indications which are used for the distinguishing between different sounds of speech. In ultrasonic components there are found new indications and some new possibilities appear for automatic speech recognition techniques. The existance of new indications has an importance for disguishing phonemes from those, which spectra in the audible frequency range are similar.

It is necessary of course to add a new ultrasonic channel to ordinary channels of audible frequency range. The automatic speech recognizer must be supplied with a set of ultrasonic band-pass filters, corresponding to different ranges of ultrasonics, namely to low, medium and high-frequency parts of the frequency range 20-160 kc/s. The spectra show the intensity distributions which allow to form suitable spectral differences used in recongnition processes.

Ultrasonic speech components may be used for segmentation of speech with many applications in speech communication systems, in speech recording and in acoustical measurements.

References

1. Forgie T.W., Forgie C.D. Progress for Recognizing the English Fricative consonants (f) and (θ), Fourth International Congress on Acoustics, Coppenhagen vol. 2 1962.
2. Miasnikov L.L. Automatic Speech Recognition, Journ. Techn. Phys. 13, 1943, 109 (in Russian).
3. Gromov M.P., Miasnikov L.L., Miasnikova H.N., Ultrasound in Speech (a Note), Soviet Acoustical Journal 13, 4, 1967 627.

THE 6TH INTERNATIONAL CONGRESS ON ACOUSTICS
TOKYO, JAPAN, AUGUST 21-28, 1968

P-0-35

Acoustical Problems in Non-Verbal Communication

Burris-Meyer, Harold

Professor, College of Humanities

Florida Atlantic University

Definition

Non-verbal communication in the context here presented is an art. It is the communication of emotion as opposed to the communication of information. The end-point of non-verbal communication is to involve the subject emotionally to the point where he acts on the basis of the emotion stimulated or communicated. The stimulated action may consist in physically or intellectually doing something which might not otherwise have been undertaken; or in the adoption of an attitude or position against which facts or phenomena are judged or evaluated.

Carried to its logical and simplest conclusion, non-verbal communication can be the source of motivation. The understanding of non-verbal communication can be the salient factor in explaining why we do the things we do.

History

The history of acoustics in non-verbal communication is long and is documented in ancient Greek and Biblical literature.[1] But lack of controls of the stimulus or definite measurement of induced response has given an aura of folk-lore to the ancient reports, and has encouraged two popular but unjustified assumptions. First: that induced action through the control of the emotional determinants thereof is a nonsensical concept because of the absence of techniques for precise definition of the stimulus and measurement of the response of the subject. Second: that music,

defined only in musical terms, when used in hospitals can induce predictable and beneficial responses in subjects with such reliability as to constitute, in fact, a therapy.[2] The first attitude has resulted in an unwillingness on the part of acoustical and psychological researchers to take the subject seriously; the second has been accompanied by a wave of charlatanism, which further inhibits serious consideration.

However, a few attempts to achieve communication of emotion by auditory means have involved sufficiently precisely defined stimuli, exact repetition, careful observation and, to some extent, measurement of the subject to indicate that further studies involving precision sufficient to establish truth can be undertaken profitably. Prominent among these indicators are: the observed audience response to the Brainstorm in Elmer Rice's ADDING MACHINE.[3] This has been extensively reported. It resulted in considerable anger induced by a 32-second, largely auditory presentation. And there was functional music, in which auditory stimulus was given much more precise definition than was usual in describing music, and in which the emotional involvement of the subject was measured in terms of his performance while at work. Functional music[4] was the subject of careful study for some years and was able to induce a change in the performance of people at work, predictable within a few percent.

The BJ spectra[5] of non-meaningful and non-associative sound induced reportable and observable nervous tension. An auditory presentation accomplished rough screening of Naval personnel for indications of nervous instability.[6]

The Problems

Only since the electronic generation and measurement of sound has it been possible to define an auditory phenomenon with great specificity, reproduce it precisely, and relate it mathematically to a measurable effect. Possible it is, but the difficulties are considerable. An auditory stimulus expressible in an art form is extremely complex, involving not only the spectrum but the intensity and time. The best we have been able to do so far is to devise a normal-appearing presentation which involves phenomena such as startle or flicker, which have been shown in the

laboratory to have identifiable effects, measure the critical elements, record the program and reproduce it under controlled conditions.

In addition to the problem on quantifying the stimulus, there is the problem of getting an unweighted response from the subject. If we control the environment by putting the subject in a calorimeter or anechoic chamber, he knows that he is a subject, and this knowledge introduces an indeterminate variable into his response, which makes any measurements taken under such tests valueless. It is therefore necessary to measure the response of the subject without his knowledge of the fact that he is a subject.

One more problem lies in the fact that subjects may not all respond uniformly, and it can be argued that results obtained for one individual may not apply to another. Though evidence to support this theory is not convincing, there is a potential problem which must be solved or circumvented.

Finally, if we are to improve the effectiveness of communicating emotion, we must relate the measured response to the measured stimulus to the end that we can prescribe a mathematically defined auditory phenomenon and achieve a predictable, uniform, measurable response when the subject is exposed to it. This will, in effect, be deriving a law.

Program

The project now in process undertakes to gain sufficient understanding of non-verbal communication to make its technically planned application feasible and the emotional involvement of the subject, predictable. As a starting point for our efforts to quantify the auditory stimulus, we have a small body of psychological literature, reports from indicative historical episodes, dance notation, and a number of mathematical approaches to music, such as those evidenced by computer composition and, of course, the Schillinger method of musical composition. Also, since response to an auditory stimulus may be influenced by what is seen, it is necessary similarly to quantify visual stimuli, and that endeavor is also under way. Though it is not subject to this paper, it should be noted that, when the visual stimulus is measured, it should be possible to effect a reconciliation between visual and

auditory stimuli to the end that we shall know how stimuli in the two media can reinforce, counterpoint, or cancel each other.

Finally, we are undertaking to measure emotional involvement: emotional state, change and rate of change, in terms of the physiological concomitants thereof. Physiological indices of emotional condition are numerous. We have determined first that we must measure the individual without his knowing he is being measured, and this rules out all devices attached to the individual or all test environments which are patently test environments. The subject must be measured as he sits in the theatre or listens to the record player at home, or watches the television or operates a lathe in a factory. Some physiological indices are remotely detectable and measurable. Measurements reduced to graphic form give a distinctive signature related to degree of emotional involvement.

Limitations and Implications

Anticipating the objection that different individuals respond differently to similar presentations, we direct our presentation and test to mass media and people en masse. Also, we recognize that it is vain to anticipate the early development of techniques that will be universally effective. But even a small percentage of gain in the emotional impact of a message can win an election. A law relating stimulus to response can help the artist to learn his art more surely and to sing his song to more men. And the teacher can devise a scheme for learning for any age and place which carries its own built-in motivation. Perhaps we may even hope that man may yet win in the race between education and catastrophe.

1. JASA, V. 17, No. 3, p. 232
2. Soibelman, Doris, Therapeutic and Industrial Uses of Music. Columbia University Press, N. Y. 1948
3. JASA, V. 11, p. 348
4. Mechanical Engineering, Jan. 1943
5. JASA, V. 32, No. 12, p. 1570
6. Ibid., p. 1572