[13] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Amer.*, vol. 27, pp. 338–352, 1955.
[14] J. M. Pickett, "Perception of vowels heard in noises of various spectra," *J. Acoust. Soc. Amer.*, vol. 29, pp. 613–620, May 1957.
[15] M. D. Wang and R. C. Bilger, "Consonant confusions in noise: A study of perceptual features," *J. Acoust. Soc. Amer.*, vol. 54, no. 5, pp. 1248–1266, 1973.
[16] R. N. Shepard, "Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space," *Psychometrica*, vol. 22, pp. 325–345, Dec. 1957.
[17] J. M. Pickett, "Perception of compound consonants," *Language and Speech*, vol. 1, pp. 288-304, 1958.
[18] D. R. Hill, "An analysis of audibility and perceptual confusion for forty English words," Loughborough College Advanced Technol., Loughborough, England, Proj. Rep., Oct. 1963.
[19] —, "An ESOTerIC approach to some problems in automatic speech recognition," *Int. J. Man-Machine Studies*, vol. 1, pp. 101–121, 1969.
[20] R. Jakobson, C. G. M. Fant, and M. Halle, *Preliminaries to Speech Analysis, the Distinctive Features and Their Correlates.* Cambridge; MA: M.I.T. Press, 1951.
[21] J. B. Kruskal, "Multidimensional scaling by optimising goodness of fit to a nonmetric basis," *Psychometrica*, vol. 29, pp. 1–27, 1964.
[22] M. B. Herscher and R. B. Cox, "An adaptive isolated-word speech recognition system," in *Proc. IEEE Conf. Speech Communication and Processing*, pp. 89–92, Apr. 1972.
[23] G. L. Clapper, "Machine looks, learns, listens," *Electronics*, pp. 91–102, Oct. 1967.
[24] V. M. Velichko and N. G. Zagoruyko, "Automatic recognition of 200 words," *Int. J. Man-Machine Studies*, vol. 2, pp. 223–234, 1970.
[25] R. J. Niederjohn and I. B. Thomas, "Computer recognition of the continuant phonemes in connected English speech," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 526–534, Dec. 1973.
[26] W. Bezdel, "Some problems in man-machine communications using speech," *Int. J. Man-Machine Studies*, vol. 2, 1970.
[27] R. F. Purton, "Speech recognition using autocorrelation analysis," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 235–239, Apr. 1968.
[28] R. K. Moore, "A simplified speech recognition system," Univ. of Essex, Colchester, England, 1973, unpublished.
[29] J. W. Forgie and C. D. Forgie, "Results obtained from a vowel recognition computer program," *J. Acoust. Soc. Amer.*, vol. 31, pp. 1480–1489, Nov. 1959.
[30] J. H. Warren, "A pattern classification technique for speech recognition," *IEEE Trans. Audio Electroacoust.*, vol. AU-19, pp. 281–285, Dec. 1971.
[31] P. Denes and M. V. Mathews, "Spoken digit recognition using time-frequency pattern matching," *J. Acoust. Soc. Amer.*, vol. 32, pp. 1450–1455, Nov. 1960.
[32] W. Bezdel, "Speech recognition using zero-crossing measurements and sequence information," *Proc. Inst. Elec. Eng.*, vol. 116, pp. 617–623, Apr. 1969.
[33] J. D. Hood and J. P. Poole, "Speech audiometry in conductive and sensoraneural hearing loss," *Sound*, vol. 5 pp. 30–38, 1971.
[34] P. W. Stevenson, "Reaction time measurements in speech discrimination tasks—An automated system with closed response sets," *J. Phonetics*, vol. 1, pp. 347–367, 1973.

Normalization of Vowels by Vocal-Tract Length and Its Application to Vowel Identification

HISASHI WAKITA

Abstract—**A new approach to speech parameter normalization is presented in which no prior knowledge about the input speakers is required. The vocal-tract length and area function are first estimated from the acoustic speech waveform, and then the area function is normalized to an acoustic tube of the same shape having a certain reference length. The normalized formant frequencies are defined as the resonance frequencies of this acoustic tube. The distributions of unnormalized and normalized formant frequencies for 9 stationary American vowels were investigated with 14 male and 12 female speakers. Fairly compact distributions of the vowels in the normalized F_1-F_2-F_3 space were obtained. A preliminary identification test for stationary vowels based on this normalization method showed an expected average recognition rate of 84-96 percent for arbitrarily selected speakers, depending on the phonetic criteria adopted for defining "correct" identification.**

I. Introduction

In the automatic identification of speech sounds produced by arbitrary speakers, normalization of the speech parameters to eliminate inter-speaker differences is indispensable. As is well known, the positions of the formant frequencies for any given vowel vary considerably from speaker to speaker. This problem of normalization has been one of the major barriers in the study of the automatic acoustic-phonetic transformation for arbitrary speakers. Various approaches have been attempted in the past to overcome this problem, e.g., [1]-[6], but many of these required prior knowledge about the input speakers. This is somewhat undesirable for a speech recognizer designed for arbitrarily selected input speakers. Although recent advances in the area of pattern recognition may provide sophisticated categorization techniques applicable to speech parameters, it seems that a more physically appealing approach might be desirable. This paper proposes a new normalization approach along this line of thinking. The approach is based on the similarity of the anatomical structures of the vocal organs among adults, and uses vocal-tract shape and vocal-tract length for normalizing speech parameters. One advantage of this approach over previous ones is that it does not require any prior knowledge about individual speakers. Another advantage is that since the estimation of the instantaneous vocal-tract length will be shown to be possible,

Manuscript received April 4, 1976; revised November 29, 1976. This work was partially supported by the Office of Naval Research under Contract N00014-67-C-0018.

The author is with the Speech Communications Research Laboratory, Santa Barbara, CA 93109.

the approach can take the length variation from one vowel to another into consideration. In fact, this is essential, since the vocal-tract length of an adult male can vary from 16.5 cm to 19.5 cm, depending on the vowel produced [7]. Although the normalization method based on the average fundamental frequency as reported by Schwartz [4] does not require prior knowledge about the input speaker, the method lacks the ability to estimate vowel-dependent lengths. Recent technological advances in speech analysis have made it possible to estimate the vocal-tract shape [8] and to estimate the vocal-tract length directly from acoustic speech waveforms.

Since the normalization approach developed in this paper is related to the articulatory parameters (specifically vocal-tract shape and length), it is believed to provide a better insight toward more general definitions of vowel sounds in the area of acoustic phonetics.

Following the description of the normalization approach and the experiment with stationary vowels, a vowel identification experiment based on the normalization approach will be described. Lastly, the method for estimating the vocal-tract length is discussed, since the method plays an important role in the normalization.

II. Normalization

A. Hypothesis

Consider a situation where a number of speakers intend to produce a stationary vowel X in a certain consonantal environment. In this case, the vowel is considered to implicitly represent some corresponding phoneme /X/. Under this circumstance, it is hypothesized that the vocal-tract configurations of the speakers are similar to each other and differ only in length. Strictly speaking, however, the vocal-tract geometry is different among men, women, and children. It is known, for example, that women and children have shorter pharynges in relation to their oral cavities [9]. Even so, among adult male and female speakers, the above assumption is not unreasonable as a first step toward inter-speaker normalization in consideration of the structural similarity of the human vocal organs from individual to individual. Consequently, it is hypothesized that if all the vocal-tract shapes are normalized to a certain reference length without altering the shapes, a statistically compact set of vocal-tract shapes for each vowel will result. A statistically compact distribution of formant frequencies would then also result after this normalization.

B. Method and Experiment

The experimental procedures used are similar to those of the Peterson and Barney study [10] except that in this study, transcription of vowels by a trained phonetician was conducted instead of large-scale listening tests as in their study. A list of words was presented to a speaker and his utterances of the words were recorded on tape. The list contained ten monosyllabic words, each beginning with /h/ and ending with /d/, and differing only in the intervening vowel, as shown in Table I. Subjects were instructed to produce these words as carefully as possible. No master sounds were presented to them. Two word lists were prepared, each arranged in a different random order. Subjects were asked to read these lists once. The subject group consisted of 26 adult speakers, including 14 men and 12 women. Fifteen of them were native Californians and the rest were residents of California from 5 to over 15 years. The 10 vowel utterances in the second list were used for analysis, thus making the total number of samples analyzed 260.

TABLE I
Vowels Used in The Experiment

Vowel	Words
/i/	heed
/ɪ/	hid
/ɛ/	head
/æ/	had
/ɑ/	hod
/ɔ/	hawed
/ʊ/	hood
/u/	who'd
/ʌ/	hud
/ɝ/	heard

A block diagram of the analysis procedure is shown in Fig. 1. Because of our familiarity with formant frequencies, unnormalized and normalized formant frequencies are compared instead of area functions.

First, the words are digitized at a sampling frequency of 10 kHz with the frequency-band limited to 5 kHz. The linear prediction autocorrelation method is applied to these speech data with a constant frame size of 30 ms, a frame shift of 6.4 ms, a first-order backward difference of preemphasis, and a Hamming window. For each frame, 10 filter coefficients are computed and the unnormalized formant frequencies are estimated by solving for the roots of the 10th-order polynomial determined by the coefficients. By visually observing the computed formant frequencies, the most stationary portion of 42.8 ms (3 frames) of each vowel utterance was chosen for the normalization experiment.

The vocal-tract length l is estimated from acoustic data by a method which will be described in Section IV. Finally, the formant frequencies $\overline{F_i}$ of the normalized vocal-tract shape are computed by multiplying the unnormalized formant frequencies by the length factor, l/l_R, where l_R is a reference length. The estimation of the vocal-tract length and the normalization of the formant frequencies by the reference length are done independently for each analysis frame.

In this normalization procedure, the vocal-tract transfer function $H(z)$ is assumed to have the form

$$H(z) = K/(1 + \sum_{i=1}^{M} a_i z^{-i}) \equiv K/A(z), \tag{1}$$

where K is a constant, the a_i's are the predictor polynomial coefficients, and M is the degree of polynomial. M also defines

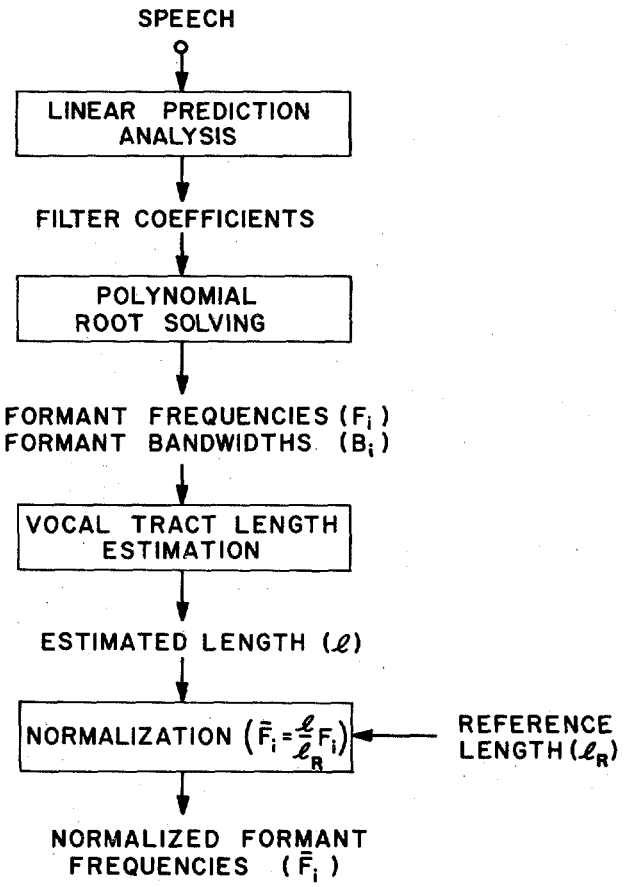

Fig. 1. Block diagram of the analysis procedure for normalization of formant frequencies by vocal-tract length.

the number of sections in the discrete acoustic tube model of the vocal tract [8]. To evaluate the frequency response of $H(z)$, z is given by

$$z = \exp(j\omega T) = \exp(j\omega 2l/Mc) \tag{2}$$

where ω is the angular frequency, T is the sampling period, and c is a sound velocity. Let the ith root of the polynomial $A(z)$ be denoted as

$$z_i = \exp\left[\frac{2l}{Mc}(-\pi B_i + j2\pi F_i)\right] \tag{3}$$

where F_i and B_i are, respectively, the ith resonance frequency and bandwidth of the vocal tract. Since $A(z)$ has roots in the form of complex conjugate pairs, i goes from 1 to $M/2$.

Based upon the hypothesis for normalization, it is necessary to compute the resonance frequencies and bandwidths of an acoustic tube when the length of the tube l is varied to a reference length l_R without altering its shape. Mathematically, if the shape of an acoustic tube is unchanged, the reflection coefficients which determine the shape of the acoustic tube are unchanged. The polynomial $A(z)$ is then unchanged because of the existence of a unique recursive relation between the reflection coefficients and the predictor polynomial coefficients. The roots z_i of $A(z)$ are also unchanged. Consequently, if the length l is changed to l_R with M and c held constant, from (3) the new ith resonance frequency $\bar{F}_i$ and bandwidth $\bar{B}_i$ must be

$$\bar{F}_i = l/l_R\, F_i \tag{4}$$

and

$$\bar{B}_i = l/l_R\, B_i. \tag{5}$$

Thus, $\bar{F}_i$ and $\bar{B}_i$ are defined as the normalized formant frequency and bandwidth, respectively. Details of estimating the vocal-tract length will be given in Section IV.

C. Results

Fig. 2 shows the distributions of unnormalized formant frequencies projected onto the F_1-F_2, F_1-F_3, and F_2-F_3 planes in an arrangement that provides an easily conceptualized three-dimensional view. The formant frequencies plotted in the figure were determined by averaging the values over the three analysis frames for the most stationary portions of the vowels. For each vowel category, the principal axes and the standard deviation along each principal axis were computed. The phoneme boundary for each vowel in the figure is defined as the 2-sigma-radius ellipse. The elliptic domain shown for each vowel was first computed using the original linear frequency scales and was then plotted on the linear-log and log-log scales used in the figure.

The vowel / ɝ / as in *heard* is not plotted in the F_1-F_2 plane to avoid confusion. The vowel /ɔ/ as in *hawed* is also omitted since it tends to merge into /ɑ/ among native Californians.

Fig. 3 shows the distributions of normalized formant frequencies. The unnormalized formant frequencies for each of the three analysis frames were normalized by use of (4) and then averaged to represent the normalized formant frequency estimates used to characterize the input vowel. The phonetic boundaries given by solid lines are again the 2-sigma-radius ellipses.

The phoneme boundaries for unnormalized vowels are also given by broken lines in the $\bar{F}_1$-$\bar{F}_2$ plane for comparison. It is seen from Fig. 3 that a fairly compact and well-separated clustering of the vowels takes place in the $\bar{F}_1$-$\bar{F}_2$ plane. It should also be noted that the overlap among /ɪ/, /ɛ/, /ʌ/, and /ɑ/ in the unnormalized F_1-F_3 plane is almost completely eliminated after normalization. Another important aspect to be noted is the direction and the amount of contraction of some of the distributions for vowels such as /ɪ/, /ɛ/, /æ/, /ʌ/, /ʊ/, and /u/. Considering the nature of the normalization process, one might assume that two adjacent clusters (e.g., for /ɪ/ and /ɛ/) could be made more compact by normalization, but that their discriminability would not be improved since the normalization only moves points along straight lines through the origin. It turns out, however, that a slight difference in vocal-tract lengths causes slightly different contractions as can be clearly seen for the vowels /ɪ/ and /ɛ/, thus improving discriminability. Similar shrinkage can also be seen in such pairs as /æ,ʌ/, /ʌ,ɑ/, and /u,ʊ/. The means, standard deviations along formant frequency axes and principal axes, and eigenvector components for each principal axis in the three-dimensional F_1-F_2-F_3 and $\bar{F}_1$-$\bar{F}_2$-$\bar{F}_3$ spaces are listed in Table II. From Table II, it is obvious that the distributions along each formant frequency axis become compact after normalization. As a measure for testing the compactness and separability

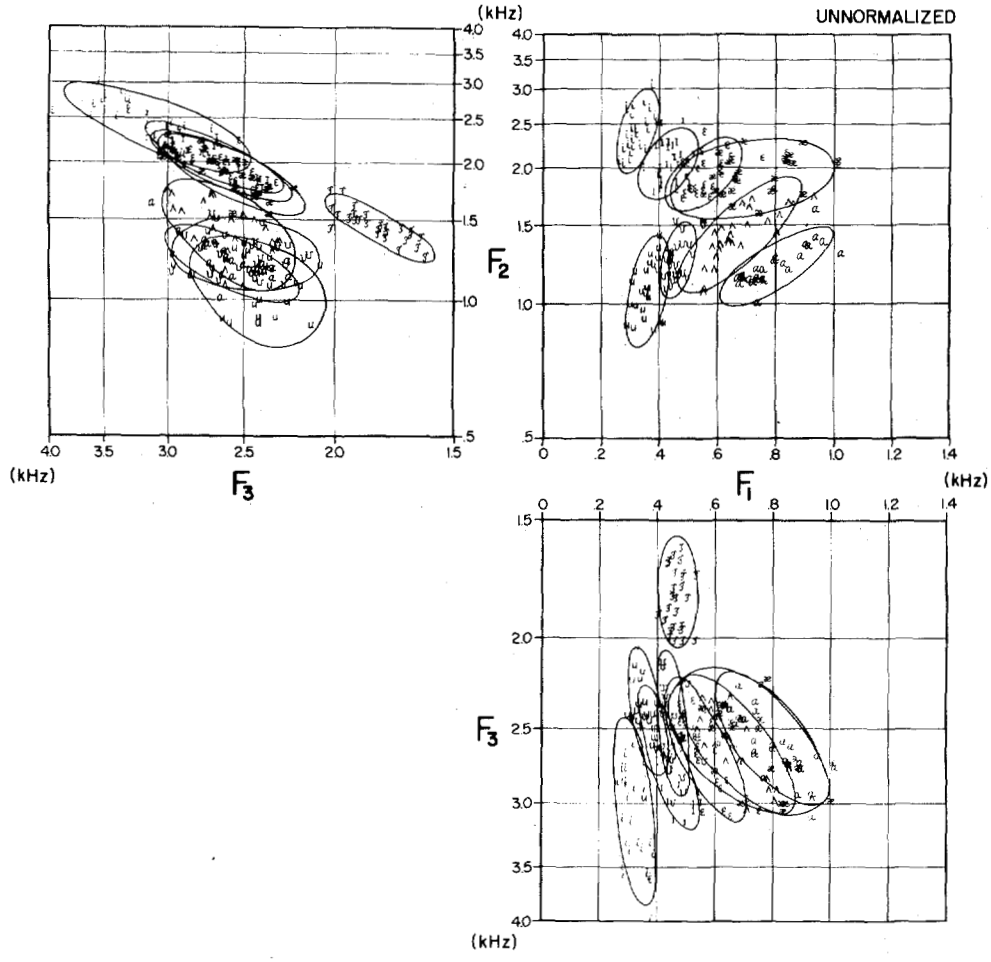

Fig. 2. Distribution of unnormalized formant frequencies projected onto the F_1–F_2, F_1–F_3, and F_2–F_3 planes. (Phoneme boundaries are defined by the 2-sigma-radius ellipses.)

of the vowel categories after normalization, the ratio of the variance between vowel categories to the variance within vowel categories was computed for pairs of closely adjacent vowel categories before and after normalization. The results are shown in Table III. The increased ratios for the normalized vowels indicate the increased separation of those vowel categories affected by normalization.

The compactness of the distributions resulting from normalization indicates that the original hypothesis is reasonably valid. Most of the histograms taken along each formant frequency axis were single-peaked, with male and female frequencies mixed. Only the $\bar{F}_1$ histogram for the /æ/ was double-peaked with each peak roughly representing each sex. The reason for this is not clear at this moment, since various factors could contribute to this. The most likely reason is the difference in vocal-tract dimensions between the sexes. Further investigation is needed on this point as more samples are collected. In contrast to the distributions for $\bar{F}_1$ and $\bar{F}_2$, the distributions along the $\bar{F}_3$ axis tended to be broad, and the $\bar{F}_3$ histograms for /u/, /ʊ/, and /ʌ/ were double-peaked. This seems to be due to the fact that the estimation of F_3 is less consistent than that of F_1 and F_2 among some speakers. This was more the case among female speakers. Taking the vowel /u/, for instance, the maximum standard deviations of male speakers for inter-frame variability of the first three formant frequencies were 16 Hz for F_1, 25 Hz for F_2, and 50 Hz for F_3 over the duration of 87.6 ms (10 frames). On the other hand, those of female speakers were 20 Hz for F_1, 30 Hz for F_2, and 298 Hz for F_3. This large fluctuation of F_3 estimates caused errors in the length estimation, and thus contributed to the broader distribution along the $\bar{F}_3$ axis. Although many of the speakers gave fairly consistent F_3 estimates, some of them gave fluctuating F_3 estimates. Why this should be so among particular speakers and how those large fluctuations can be eliminated remain as questions to be answered.

III. Identification of Normalized Vowels

Since our main interest was to apply the normalization method to automatic sound identification, a recognition experiment was conducted to test the applicability of the method. This recognition experiment was conducted with 10 new speakers, when analysis of vowels from the first 16 speakers (9 males and 7 females) was completed.

The same analysis procedure shown in Fig. 1 is used to compute the normalized formant frequencies and then a simple Bayesian decision rule is applied for identification of the input sound. Since the normalization scheme isolates the vowels fairly well, no sophisticated pattern recognition

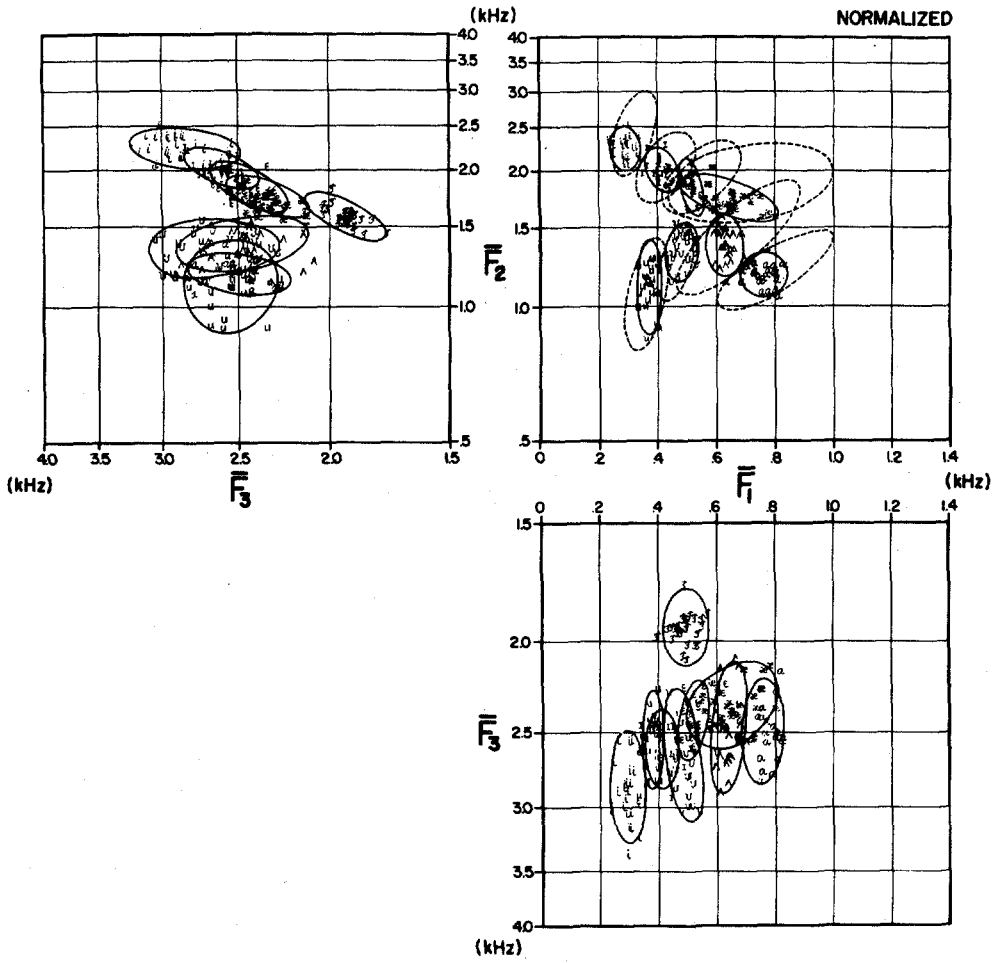

Fig. 3. Distribution of normalized formant frequencies projected onto the $\bar{F}_1$-$\bar{F}_2$, $\bar{F}_1$-$\bar{F}_3$, and $\bar{F}_2$-$\bar{F}_3$ planes. (Phoneme boundaries are defined by the 2-sigma-radius ellipses.)

techniques were considered for this experiment. By studying the histograms of formant frequencies along each frequency axis, normal distribution was assumed for the first three normalized formant frequencies for each vowel. The likelihood of the first three normalized formant frequencies of the input vowel frame was computed for each reference vowel using this assumption. The likelihood $L_i(X)$ that the feature vector X of the input signal belongs to the ith reference vowel X_i is given by

$$L_i(X) = \ln p_i - \tfrac{1}{2}(X-\mu_i)^T V_i^{-1}(X-\mu_i) - \ln\{(2\pi)^{3/2}|V_i|^{1/2}\} \quad (6)$$

where p_i is an *a priori* probability of occurrence for the vowel X_i, μ_i is the mean vector for X_i, V_i is the covariance matrix for X_i, and T denotes the transpose of a matrix [11]. In this experiment, equal *a priori* probabilities $p_i = \frac{1}{9}$ were assigned. As reference data, μ_i and V_i for each vowel were computed from the normalized formant frequencies of the 16 reference speakers. The input frame was identified as the phoneme corresponding to the class which gave the maximum likelihood in (6). In the recognition experiment, the total number of vowels identified was 90. The linear prediction method was applied for analysis. The same analysis conditions as used in the previous section were also used in this experiment. For identification of vowel categories, the most stationary 55.6 ms (5 frames) of the utterances in which steady-state portions ranged approximately from 60 to 90 ms were manually chosen by observing the first 3 unnormalized formant frequencies.

In this identification task, it was assumed that each speaker intended to produce the canonical vowels and that these target vowels were realized. Under this assumption, the overall rate of correctly identifying the vowels was 84.4 percent. In making the decision, a correct score was given to those vowel segments in which three or more frames out of the five were labeled correctly.

A similar experiment for unnormalized vowels resulted in 78.9 percent. For normalized vowels, higher scores were obtained for all the vowels except for /i/ and /u/ for which the scores were unchanged after normalization.

The details of the recognition score are shown in Fig. 4. The score of each speaker is shown in each row. The speakers indicated by M are male speakers, and those indicated by F are females. The circles show correct scores. The scores for individual speakers are shown in the right-most column, and the scores for the different vowels are shown in the bottom row. Mislabeled vowels are indicated in their corresponding cells.

TABLE II
MEANS, STANDARD DEVIATIONS, AND EIGENVECTORS FOR PRINCIPAL AXES COMPUTED FOR THE UNNORMALIZED AND NORMALIZED FORMANT FREQUENCIES FOR 9 AMERICAN VOWELS PRODUCED BY 26 SPEAKERS (14 MALES AND 12 FEMALES). (U: UNNORMALIZED, N: NORMALIZED)

		Mean (Hz)			Standard Deviation along formant axis (Hz)			Standard Deviation Along Principal Axis (Hz)	Eigenvector		
		F_1	F_2	F_3	F_1	F_2	F_3		r_{F_1}	r_{F_2}	r_{F_3}
i	U	324	2450	3149	38.3	270.1	363.4	429.4	.039	.565	.824
								145.8	-.009	.825	-.565
								34.5	.999	-.014	-.037
	N	299	2246	2888	31.9	123.7	193.3	198.0	.007	.268	.963
								116.0	-.021	.963	-.268
								31.8	1.000	.018	-.012
ɪ	U	444	2089	2716	52.7	186.9	230.6	285.5	.125	.605	.787
								90.6	-.227	.789	-.571
								34.1	.966	.107	-.236
	N	425	2003	2604	30.2	121.0	110.8	143.5	-.029	.764	.644
								80.5	.151	-.634	.759
								27.6	.988	.120	-.097
ɛ	U	567	1984	2654	73.0	178.9	240.3	288.6	.190	.551	.813
								98.5	-.021	.831	-.514
								45.3	.959	.073	-.274
	N	521	1833	2448	26.3	127.1	119.2	152.7	-.080	.748	.659
								84.9	.055	-.657	.752
								23.0	.995	.096	.012
æ	U	723	1934	2615	142.1	199.5	239.5	307.2	.268	.593	.759
								127.2	.870	-.487	.073
								82.2	.413	.641	-.646
	N	650	1732	2356	92.9	140.1	167.9	202.6	-.104	.614	.783
								103.1	.768	-.451	.455
								69.2	.632	.648	-.425
ɑ	U	798	1233	2594	102.6	122.7	214.6	245.3	.340	.392	.855
								89.3	.333	.800	-.499
								58.9	.879	-.455	-.142
	N	767	1187	2502	48.4	71.5	165.2	166.2	.027	.111	.993
								69.6	-.085	.990	-.109
								48.1	.996	.082	-.037
ɔ	U	462	1271	2498	33.1	123.7	224.0	224.7	.056	.069	.996
								123.5	.085	.994	-.074
								29.0	.995	-.089	-.050
	N	495	1359	2675	32.6	106.0	226.2	227.3	.044	-.102	.994
								104.1	.045	.994	.100
								30.7	.998	-.040	-.049
ʊ	U	363	1102	2430	35.3	157.2	196.1	210.7	.109	.467	.878
								139.0	.007	.882	-.470
								27.0	.994	-.057	-.092
	N	379	1153	2544	19.1	144.1	150.0	150.2	-.011	-.151	.988
								144.0	.003	.989	.151
								19.1	1.000	-.005	.010
ʌ	U	677	1452	2613	108.3	202.7	222.5	282.4	.342	.616	.709
								141.5	.104	.725	-.681
								50.0	.934	-.307	-.184
	N	641	1380	2501	33.0	98.8	205.9	207.7	-.041	-.138	.990
								95.5	-.048	.990	.136
								31.6	.998	.042	.047
ɝ	U	464	1491	1815	29.9	130.1	134.5	177.9	-.023	.693	.721
								60.0	-.264	.691	-.673
								25.9	.964	.206	-.167
	N	499	1596	1945	42.4	87.7	95.7	114.2	-.105	.654	.749
								67.3	-.411	.657	-.632
								33.0	.906	.374	-.200

A further scrutiny of the phonetic transcription by a trained phonetician revealed that the errors were categorized into four types. Type I error is an apparent mispronunciation. Type II error is caused by allophonic variation which deviated considerably from the defined target vowels. This type II error tends to reflect the speakers' dialect characteristics. Type III error is due to either inaccurate estimation of formant frequencies and bandwidths, or to inaccurate length estimation. Type IV error is an error due to unknown factors. These types of errors are represented by the corresponding numerals in Fig. 4. No correctly identified vowels were mispronounced or produced with noncanonical allophones.

In this experiment, the occurrence of the type I and II errors was relatively high, particularly among those subjects who were not accustomed to the recording situation. Type III errors were relatively infrequent, and only one case of type IV error is seen in this experiment. The vowel /æ/ of speaker 8 was categorized as /ɛ/. The unnormalized first and second formant frequencies, in this case, fell in the middle of the /ɛ/ region in the Peterson-Barney phoneme boundary, although it was correctly transcribed as /æ/ by the phonetician. The reason for this is not clear. The third formant is probably influential, although it fell between those expected for /æ/ and /ɛ/.

Taking these types of errors into account, the recognition rate of correct identification is improved to 90.5 percent if

TABLE III
THE RATIOS OF VARIANCE BETWEEN GROUPS TO VARIANCE WITHIN GROUPS FOR VARIOUS COMPONENTS COMPUTED FOR PAIRS OF ADJACENT VOWELS. (U: UNNORMALIZED, N: NORMALIZED)

		The ratio of variances between and within groups				
		F_1	F_2	F_3	F_1-F_2	F_1-F_2-F_3
i ↔ ɪ	U	1.763	.626	.525	.669	.579
	N	4.334	1.027	.844	1.227	.994
ɪ ↔ ɛ	U	.960	.085	.018	.180	.083
	N	2.970	.492	.474	.614	.551
ɛ ↔ æ	U	.498	.018	.007	.144	.070
	N	1.172	.191	.142	.401	.278
æ ↔ ɑ	U	.096	4.670	.002	3.028	1.372
	N	.770	9.496	.251	6.366	2.475
æ ↔ ʌ	U	.035	1.496	.000	1.083	.556
	N	.006	3.082	.185	2.243	.855
ɑ ↔ ʌ	U	.347	.446	.002	.418	.189
	N	2.385	1.304	.000	1.507	.313
ʊ ↔ u	U	2.183	.370	.027	.470	.170
	N	4.879	.694	.121	.873	.356
ʊ ↔ ʌ	U	1.860	.304	.069	.592	.283
	N	5.149	.011	.168	.488	.232

type I error is neglected, and to 96.2 percent if both type I and II errors are neglected. Based on the investigation of the results of acoustical analysis of the first 16 speakers used as reference, plus the 10 new speakers, the anticipated recognition rate ranges between 84 and 96 percent, depending on the particular definition of what constitutes an error. Although there is a noticeable difference in the recognition rate between male and female speakers from this experiment, the actual difference seems to be slight. The difference seems to be rather speaker-dependent. This indicates that a high recognition rate might be maintained for a limited number of speakers by a careful selection of the speakers.

IV. VOCAL-TRACT LENGTH ESTIMATION

As can be seen from the previous sections, it is indispensable in this normalization method to be able to estimate the vocal-tract length simultaneously with other speech parameters, since the tract length varies from sound to sound. Consequently, a method for estimating the vocal-tract length from speech waveforms is discussed in this section.

One idea that has been suggested for computing the length is to use the higher formant frequencies since they tend to be regularly spaced. Assuming that those higher formant frequencies do not deviate much from those of a uniform tube having the same length, the length l is estimated from the ith formant frequency F_i as

$$l = \frac{(2i-1)c}{4F_i}. \tag{7}$$

Since the accuracy of this approach was not known, it was evaluated by a simple experiment.

Fig. 5 shows area functions for various vowel-like configurations with the vocal-tract length fixed at 17 cm. The shapes for the vowel-like area functions are identical to those obtained by Fant [9], except for their lengths. On the left-hand side, their corresponding frequency spectra are shown. The bottom tube is a uniform tube for which the resonance frequencies are uniformly spaced. But, for the vowel-like configurations which deviate strongly from the uniform one,

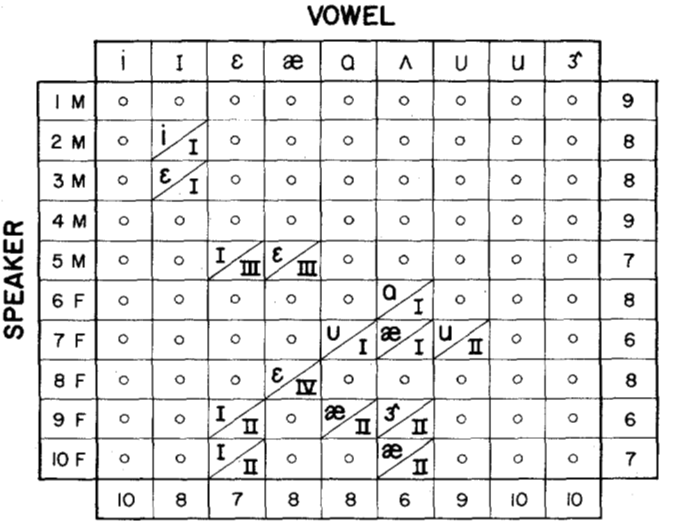

Fig. 4. Results of the identification experiment on normalized steady-state vowels. The symbols designate mislabeled sounds and the Roman numerals indicate the types of the errors.

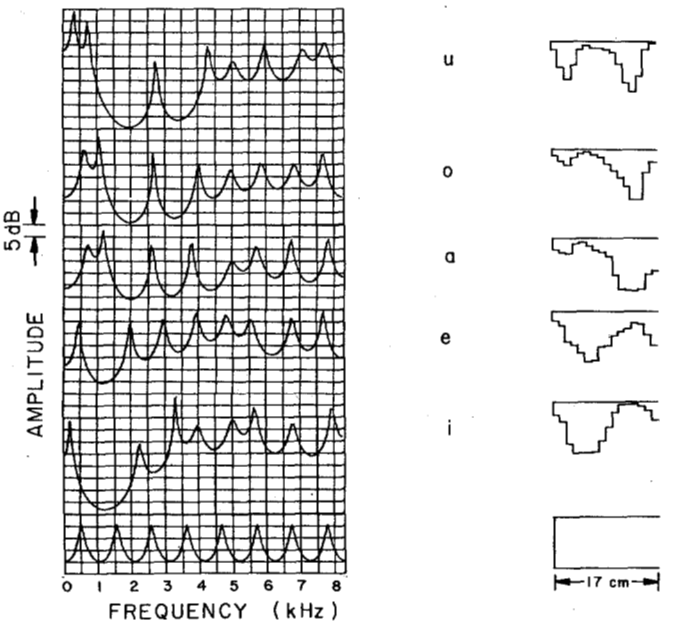

Fig. 5. Vowel-like acoustic tube shapes and their resonance characteristics.

the higher resonance frequencies are not as regularly spaced as might be expected.

The use of only the fourth resonance frequency for tube lengths of 19, 17, and 15 cm resulted in errors in the length estimation ranging from 4.4 to 15.6 percent. By taking the average of the lengths estimated from the fourth and fifth resonance frequencies, the error was improved, ranging from 5.2 to 11.3 percent. The use of average length estimated from the fourth through the eighth resonance frequencies improved the error further, ranging from 1.4 to 6.3 percent. It would be difficult in practice, however, to obtain higher formant frequencies for estimating the length of the vocal tract, especially when the length information is continually needed in processing connected speech.

Paige and Zue [12] have proposed a different method for computing vocal-tract length by use of an approximate relation between the pole and zero frequencies of the lip impedance and the Fourier cosine expansion coefficients of the logarithm of an area function. By applying this approximate relation to the first few pairs of pole and zero frequencies, the length was determined by minimizing a certain error criterion.

Their results obtained on Russian vowels showed a constant negative bias between the measured values and the computed ones. This result seemed to be improved by either introducing a certain ad hoc correction factor or by increasing the number of poles and zeros. However, the difference was found to be essentially due to the approximate relationship between the pole and zero frequencies and the Fourier expansion coefficients. Thus, the method is not sufficiently accurate as long as the approximation is involved. As will be shown below, however, the error criterion used in their method gives quite reasonable estimation of vocal-tract lengths if it is based on formant frequencies and bandwidths.

As a third choice, the use of formant frequencies together with their corresponding bandwidths is considered. As is well known, an infinite number of shapes of different lengths are realizable for a given set of formant frequencies and bandwidths. Among those shapes, the length corresponding to a shape which gives the minimum of the following error criterion is chosen as the actual length. The error criterion is represented as

$$\epsilon(l) = \frac{1}{M} \sum_{i=1}^{M} [\ln S_i]^2 \tag{8}$$

under the constraint

$$\sum_{i=1}^{M} \ln S_i = 0 \tag{9}$$

where S_i is the ith cross-sectional area of a vocal-tract area function, and M is the number of sections used to represent the area function. In this case, the vocal-tract shape is represented by a concatenation of cylindrical sections of equal length. The above error criterion is the discrete case of the one hypothesized by Paige and Zue [12]. As they pointed out, (11) implies that the area function is normalized so that the error becomes zero for a uniform shape. Thus, this criterion is a measure of the uniformity of the shape.

It should be noted that there is as yet no known theoretical reason why this criterion should give a reasonable estimation of the vocal-tract length, even though our experimental results support the above hypothesis.

The algorithm to determine the vocal-tract length from the acoustic speech waveforms is as follows.

1) The polynomial coefficients of the vocal-tract transfer function in (1) are computed from input speech samples by the linear prediction method.

2) By solving for the roots of $A(z)$ in (1), formant frequencies and bandwidths are computed from (3).

3) Suppose the first N formant frequencies and bandwidths are used. For a given length $l = l_m$, z_i's for $i = 1,2,\cdots,N$ are computed by use of (3) from given formant frequencies and bandwidths.

4) From the z_i's, $2N$ predictor coefficients of $A(z)$ are computed via

$$A(z) = \prod_{i=1}^{N} (1 - z_i z^{-1})(1 - \bar{z}_i z^{-1})$$

where $\bar{z}_i$ is the complex conjugate of z_i.

5) From the $2N$ predictor coefficients, $2N$ reflection coefficients are computed by use of a well known recursive relation [8].

6) The area function A_i, $i = 1,2,\cdots,2N$, is computed from the reflection coefficients [8]. Then

$$\ln S_i = \ln A_i - \frac{1}{2N} \sum_{i=1}^{2N} \ln A_i \qquad i = 1, 2, \cdots, 2N$$

is chosen to satisfy (9).

7) The error $\epsilon(l_m)$ is computed from S_i's for $M = 2N$ by use of (8).

8) By repeating the steps (3)–(7), $\epsilon(l_m)$ is computed for $m = 1, 2, \cdots, p$.

9) The length l_m which gives the minimum value of $\epsilon(l_m)$ is chosen as the value of the estimated vocal-tract length.

In the above algorithm, the range of length, for example, may be chosen from 10 to 22 cm. However, it should be noted that the longest length l_H is constrained by the location of the highest formant frequency F_H via $l_H = Mc/4F_H$. The reason is that $l > l_H$ violates the constraint of a fixed number of formant frequencies in the frequency band under consideration. This is a constraint imposed by the analysis and does not necessarily hold in the real situation.

The error criterion was tested on Fant's area functions for five Russian vowels by utilizing (3)–(9) of the above algorithm. In this case, the original area functions were converted into eight-section representations, and the first four formant frequencies and bandwidths were used. The results are shown in Fig. 6. The black circle shows the measured length while the triangle shows the length obtained by this method. The error of estimating the length was less than 4.6 percent, which was within the error range of the original measurements of the length [12]. The use of only three formant frequencies and bandwidths resulted in less accuracy for the estimation of vocal-tract length for each vowel shape. Although no extensive study has been done to compare the vocal-tract lengths estimated from actual utterances by use of this method with those of direct measurements, a preliminary study using a single speaker resulted in an error ranging between 1.6 and 8.6 percent for five vowel utterances, where the comparison was made with the length measured from mid-sagittal X-ray photographs. The results are shown in Table IV. To evaluate the accuracy of the method, a more extensive study will be needed. It should also be pointed out that bandwidth estimation is a crucial factor in this method for obtaining a better estimate of the tract length. It will probably be more desirable to be able to process the closed phase of a glottal cycle so that more accurate and consistent bandwidth information can be obtained. This remains as a problem to be pursued. The

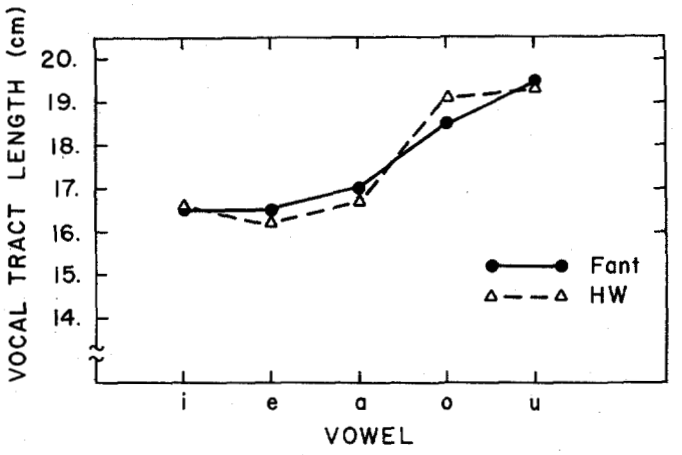

Fig. 6. Vocal-tract lengths estimated from Fant's area functions for five Russian vowels and the measured ones.

TABLE IV
MEASURED AND COMPUTED VOCAL-TRACT LENGTHS OF 5 JAPANESE VOWELS PRODUCED BY A MALE JAPANESE SPEAKER

	LENGTH		
Vowel	Measured (cm)	Computed (cm)	Error (%)
i	16.5	16.1	-2.4
e	16.2	17.6	+8.6
a	17.5	16.6	-5.1
o	18.9	19.2	+1.6
u	17.4	18.6	+6.9

TABLE V
AVERAGE ESTIMATED VOCAL-TRACT LENGTH AND STANDARD DEVIATION COMPUTED FOR 26 SPEAKERS (14 MALES AND 12 FEMALES)

	Male		Female	
	Average Length (cm)	Standard Deviation (cm)	Average Length (cm)	Standard Deviation (cm)
i	17.0	0.48	14.3	1.13
ɪ	17.4	0.81	15.2	0.83
ɛ	17.0	0.78	14.3	1.12
æ	16.3	0.88	14.5	1.32
ɑ	17.4	0.58	15.2	1.11
ʊ	19.1	1.23	17.5	1.00
u	19.1	1.20	16.8	0.82
ʌ	18.0	0.72	14.7	1.22
ɝ	19.2	1.09	17.2	1.03
Total Average	17.8	1.07	15.5	1.29

average estimated vocal-tract length for each vowel which resulted from the analysis in Section I is also shown in Table V for both male and female speakers. The overall average estimated vocal-tract length for the females is 13 percent shorter than that for the males. This result is in good agreement with the results based on X-ray measurements [9]. Furthermore, the average length for each vowel is fairly consistent with what would be expected physically. For example, due to the lip protrusion, the estimated length for the vowel /u/ is longer than that of /i/.

V. DISCUSSION

Speech sounds can be defined in three major domains: articulatory, acoustic, and perceptual. It is thus desirable to be able to define an equivalent vowel feature space in any of the three domains. In this paper, an attempt was made to utilize currently available techniques to bring the articulatory parameters of vocal-tract shape and length into the acoustic domain in order to eliminate inter-speaker differences in acoustic parameters. Thus the vocal-tract shapes of different lengths which are determined by specifying the positions of articulators according to the definition of a particular sound in the articulatory domain can be related to a particular point in the acoustic domain, i.e., a point defined by an acoustic tube having a certain reference length. In reality, sounds constituting a certain region around this point must be perceived as equivalent sounds. This region, which would be a subset of the region obtained by the current experiment, should eventually be determined. Likewise, the articulatory variations which correspond to the regions of perceptually equivalent vowels in the acoustic domain is another important question to be pursued. The normalization method described in this paper would contribute to this line of study, as well as to the automatic acoustic-phonetic transformation.

It should be noted that the normalization method discussed in this paper produces centralization of the vowel space due to the fact that all the vowels are normalized according to a constant reference length since unidentified vowels were assumed in this study. However, it will be apparent that, for known vowels, more phonetically useful normalization may be accomplished by normalizing each vowel according to the average tract length found for the category of that vowel. This will be especially desirable for the study of acoustic phonetics.

As mentioned before, the dimensions of the vocal tract vary among males, females, and children. Thus, as reported by Fant [5], [9], by taking the parameters such as the pharynx length and the oral cavity length into consideration, more compact distributions for normalized formant frequencies are expected to result, although the current acoustic analysis technique has not reached the point where the computation of such parameters from an instantaneous acoustic analysis is possible. From a perceptual point of view, however, it remains as an important question as to whether phonetically equivalent sounds come from uniformly-scalable or nonuniformly-scalable vocal-tract shapes, and whether the normalization by only the length gives sufficient discrimination of vowels.

The choice of the reference length is rather arbitrary. The

reference length of 17 cm in this study was chosen simply because it is roughly the vocal-tract length of the average adult male speaker. One could choose a reference length of 1 cm, in which case the normalized formant frequencies become the original formant frequencies multiplied by the actual estimated length in centimeters. The question of what the best specification of the normalized vowel space is is an interesting problem that is currently being investigated.

VI. Conclusion

A new approach to vowel normalization was presented, and an identification experiment on steady-state vowels indicates that the method is quite promising. This method is currently being applied to vowel identification in connected speech. Besides its application to the recognition problem, this method is expected to contribute to a definition of the vowel system of a language or dialect in the acoustic domain, irrespective of speakers with possible eventual application to the problem of automatic language identification. There are still technical problems to be improved upon. For example, problems of consistent formant frequency and bandwidth estimation and more accurate length estimation remain, in order to establish the method as a more reliable one.

Acknowledgment

The author would like to thank Dr. D. J. Broad, Dr. H. Kasuya, and Dr. J. D. Markel for their valuable discussions and comments. Thanks are also due to Dr. M. Hirano and Dr. H. Matsushita at Kurume University, Japan, for their cooperation in preparing X-ray photographs.

References

[1] H. Suzuki, H. Kasuya, and K. Kido, "The acoustic parameters for vowel recognition without distinction of speakers," presented at the 1967 Conf. Speech Communication and Processing, Bedford, MA, 1967, Paper B5.

[2] L. J. Gerstman, "Classification of self-normalized vowels," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 78–80, Mar. 1968.

[3] W. Klein, R. Plomp, and L. C. W. Pols, "Vowel spectra, vowel spaces, and vowel identification," *J. Acoust. Soc. Amer.*, vol. 48, pp. 999–1009, Apr. 1970.

[4] R. M. Schwartz, "Automatic normalization for recognition of vowels of all speakers," M.S. thesis, Massachusetts Inst. Technol., Cambridge, 1971.

[5] G. Fant, "Nonuniform vowel normalization," Speech Transmission Lab., Royal Inst. Tech., Stockholm, Sweden, Quart. Progr. Status Rep., pp. 1–19, 2–3, 1975.

[6] P. E. Nordström and B. Lindblom, "A normalization procedure for vowel formant data," presented at the Int. Congr. Phonetic Sci., Paper 212, Leeds, England, Aug. 1975.

[7] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.

[8] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 417–427, Oct. 1973; also "Estimation of the vocal tract shape by optimal inverse filtering and acoustic/articulatory conversion methods," Speech Commun. Res. Lab., Inc., Santa Barbara, CA, Mono. 9, July 1972.

[9] G. Fant, *Speech Sounds and Features*. Cambridge, MA: M.I.T. Press, 1973.

[10] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, pp. 175–184, Mar. 1952.

[11] D. F. Morrison, *Multivariate Statistical Methods*. New York: McGraw-Hill, 1967.

[12] A. Paige and V. W. Zue, "Calculation of vocal tract length," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 268–270, Sept. 1970.