Multidimensional Representation of Personal Quality of Vowels and its Acoustical Correlates

HIROSHI MATSUMOTO, SHIZUO HIKI, TOSHIO SONE, and TADAMOTO NIMURA

Abstract—The personal quality of sustained vowels uttered by eight male talkers was represented multidimensionally in a psychological auditory space (PAS) by means of Kruskal's multidimensional scaling procedure based on the perceptual confusion in talker discrimination tests. Physical properties of the vowels were analyzed in terms of elementary acoustical parameters, such as formant frequencies, slope of glottal source spectrum, mean fundamental pitch frequency, and rapid fluctuation of fundamental pitch period. Then the relationship between the configuration on the PAS and the acoustical parameters was examined through multiple correlation and regression analysis.

The contribution of those acoustical parameters to the personal quality of the five Japanese vowels and the relative contributions of the vocal tract and the glottal source characteristics are demonstrated quantitatively. These results were obtained partially by utilizing hybrid voices in which the source wave or the formant frequency pattern was interchanged among different talkers.

I. Introduction

As part of a general study investigating the auditory process for extracting personal information from speech, the relation between the perceptual difference in personal quality and the difference in physical properties was analyzed for sustained vowels.

In order to observe the perceptual difference in personal quality, recognition rates or confusion matrices have been utilized in most of the previous studies. In this study, however, it was tried to observe quantitatively the multidimensional nature that underlies personal quality in the psychological process in terms of distance on the psychological auditory space (PAS).

In the first stage of the auditory process for extracting personal information from speech, voice input is mapped onto a sensory auditory space through an elementary auditory process that deals with sensory differences in the basic attributes of sound, such as intensity, pitch, and spectral pattern. Then, we hypothesize, through higher auditory processing the sensory auditory space is mapped onto a PAS in which interpoint distance relates monotonically to the perpetual dissimilarity of personal quality of voice. (This space is independent of or parallel with the PAS of phonetic quality [1], which is also mapped from the sensory auditory space.) In the last stage, in order to output personal information, the judgment process is applied to the personal quality represented in the PAS. In the ordinary case, the process of judgment is the identification of the talker, which may be ascribed to a discrimination between the representation of a given voice input in the PAS and that of the voice characteristics of familiar talkers stored in the long-term memory.

In this experiment, the personal quality of sustained vowels was scaled multidimensionally, utilizing discrimination tests in which listeners were supposed to store the representation of the personal quality in the PAS of the preceding stimulus in short-term memory and compare it with that of the following stimulus. In this way the nature of the PAS is observed separately from the ordinary identification process. This will serve to avoid the ambiguities caused by the listener's familiarity with the talker and uncertainty of the memory.

II. Voice Samples of the Vowel /a/

Voice samples used in the first experiment were 24 sustained vowels, Japanese /a/, uttered with three levels of fundamental pitch frequency (120, 140, and 160 Hz, approximately) by each of eight male adult talkers (voice set I). The vowel /a/ was used here because this occures most frequently among the five vowels in Japanese speech.

These eight talkers were chosen as the representatives of 25 candidates in the age range from 20 to 35 years. There was no candidate with pathologic voice. The talkers were instructed to utter the vowel for a few seconds with natural intensity, adjusting its pitch frequency to that of a pure tone (120, 140, and 160 Hz) that was presented to one ear of each talker through an earphone.

Then, a 0.5-s portion of the steady part was extracted from each of these sustained vowels by reproducing the master recording through a gate circuit with a 10-ms rise and fall time. The intensity was adjusted in dubbing the submaster recordings so that the peak volume units (vu) meter reading was the same for each voice sample.

III. Determination of Acoustical Parameters

The acoustical parameters used here in examining their relation to personal quality were as follows: the lowest three formant frequencies (F_1, F_2, and F_3); the slope of glottal source spectrum (α); the mean

Manuscript received April 12, 1972; revised May 4, 1973.

The authors are with the Department of Electrical Engineering and the Research Institute of Electrical Communication, Tohoku University, Sendai, Japan.

logarithmic fundamental pitch frequency (log $\bar{F}_0$); and the rapid fluctuation of fundamental pitch period ($\sigma(\Delta T/\bar{T})$ (the standard deviation of differences between adjacent fundamental pitch periods normalized by the mean fundamental pitch period).

The formant frequencies were estimated by means of an analysis-by-synthesis method [2], applied to the log-amplitude spectrum calculated by the fast Fourier transform (FFT) from the digitized waveform of a single pitch period for each voice sample. In the analysis-by-synthesis algorithm, the six parameters (first, second, third, and fourth formant frequencies, the higher pole correction term, and the slope of glottal source spectrum in dB/octave) are automatically controlled by means of the maximum neighborhood method [3] so that the synthesized spectrum gives the best fit to the input spectrum in the frequency range from 0.1 to 4.0 Hz.

The mean value and the rapid fluctuation of fundamental pitch frequency were calculated on the basis of measurements of the fundamental pitch periods of the voice samples with an accuracy of 0.02 ms. These measurements were obtained from photographs of the speech waveforms.

In the results, the ratios of standard deviation to mean value of the first, second, and third formant frequencies for the 24 voice samples were 7.8, 5.4, and 8.0 percent, respectively, α ranged from - 6 to - 20 dB/octave and $\sigma\ (\Delta T)/\bar{T}$ from 0.4 to 1.0 percent in voice set I.

IV. Construction of the PAS

In order to measure the dissimilarity of personal quality among the voice samples, every possible pair of the voice samples (24 × 24 = 576) was presented nine times in random order to a group of six listeners. The interval between the two stimuli in a pair was 2 s. The listeners, none of whom had been familiar with the voices of the talkers, were asked to state whether or not they believed the two voice samples in a pair were uttered by the same talker and, at the same time, to indicate the degree of confidence in the correctness of their "same talker" or "different talker" judgment on a three-point scale ("very sure," "think response is correct," and "best guess"). The listeners were not given any other information about the voice samples, such as the number of talkers or the fact that the pitch frequency, intensity, and duration had been controlled.

In order to examine the homogeneity of the rate of "different talker" response of the six listeners, the correlation coefficients of the rates of "different talker" response between each of the six listeners were calculated from responses to 64 pairs sampled randomly from all 576 pairs. As the results were in the range from 0.59 to 0.75 (which was beyond the 1 percent level of confidence), it was assumed that the dissimilarity judgments of the six listeners were fairly homogeneous. Therefore, the rate of "different talker" responses were averaged for all six listeners and the nine trials. Then, the measure of dissimilarity for every possible combination of the 24 voice samples (${}_{24}C_2$ = 276 in total) was calculated by averaging these averaged responses for each of two pairs that contained nonidentical stimuli but that were themselves identical except for order, with the result that each measure of dissimilarity is based on 108 (= 6 × 9 × 2) responses.

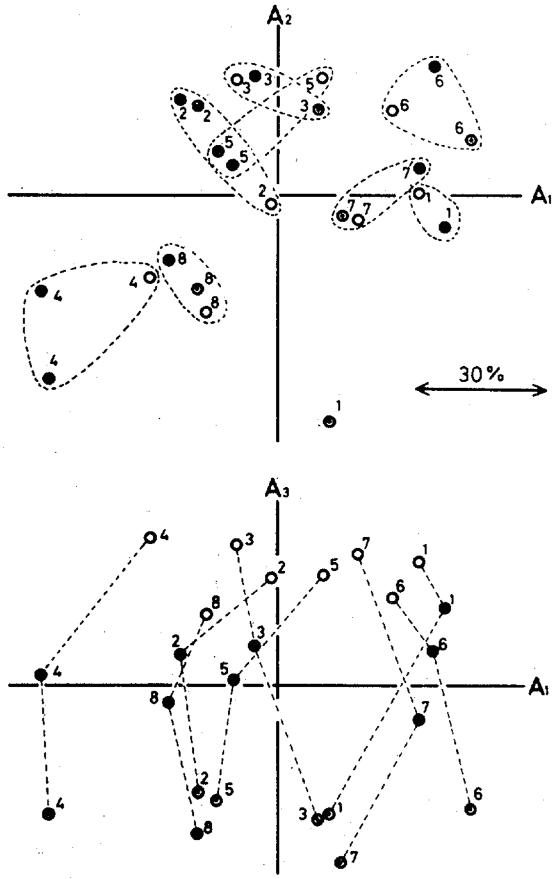

Fig. 1. Configuration on A_1-A_2 and A_1-A_3 planes of the three-dimensional PAS for voice set I, which consists of vowel /a/ uttered by eight male talkers (indicated by numerals and encircled or combined by dotted lines) with fundamental pitch frequencies of 120, 140, and 160 Hz (indicated by double, closed, and open circles, respectively). Stress is 11.5 percent. The arrows indicate the distance that approximately corresponds to 30 percent "different talker" response.

The configuration of the voice samples was derived from the measures of dissimilarity by means of Kruskal's multidimensional scaling method [4]. The configuration of voice set I is shown three-dimensionally in Fig. 1. A three-dimensional configuration was chosen since the relation to each of the acoustical parameters can be interpreted fairly reasonably on this configuration, although the stress for the three-dimensional configuration is still somewhat high and there was not a large decrease of stress from two dimensions to three dimensions (38.0, 16.5, 11.5, and 7.5 percent in one, two, three, and four dimensions, respectively).

In Fig. 1, the coordinate axes are rotated so that the A_3 axis corresponds best to the mean logarithmic fundamental pitch frequency. Then, the A_1 axis corresponds best to the slope of the glottal source spectrum on the plane perpendicular to the A_3 axis on the basis of the multiple correlations described in Section V. The "distance" is illustrated by an arrow whose length corresponds to the "different talker" response of about 30 percent.

In Fig. 1, the points that represent the sustained vowels uttered by a single talker move upward along the direction of the A_3 axis as the fundamental pitch frequency increases from 120 to 160 Hz. This indicates that the mean fundamental pitch frequency is one of the important cues for the perception of personal quality as has also been observed in a speaker identification test with vowels [5]. On the other hand, the points for a single talker cluster together in a particular region in the A_1-A_2 plane and the region for each talker covers a considerable area on the plane, showing that some other information on personal quality independent of fundamental pitch frequency is contained in these voice samples.

In order to examine this in terms of discrimination score, the receiver operating characteristic (ROC) curve was derived from the confidence ratings for each of the six combinations of the three levels of fundamental pitch frequency. The ROC curves were traced out by plotting the cumulative sum of the percent false rejection versus that of the percent correct acceptance of the six categories of similarity scale (ranging from "very sure different" to "very sure same") that were converted from the three-point confidence ratings for each of the "same talker" and "different talker" judgments. The average results for the six listeners are shown in Fig. 2 together with percent correct score P(c), a criterion-free index, which the listener would have obtained had he used a criterion for making a binary decision such that percent correct acceptance is equal to percent correct rejection. Fig. 2 indicates that the two voice samples uttered by a talker with different pitch frequencies of 160-140 Hz, 140-120 Hz, and 160-120 Hz were discriminated correctly with P(c) of 67, 60, and 56 percent, respectively.

The contribution of the same amount of change in fundamental pitch frequency to perception of the personal quality of voice becomes larger as the frequency becomes lower. This is shown by the fact that the stimulus points with 140 Hz fundamental pitch frequency are closer on the PAS in Fig. 1 to those with 160 Hz than to those with 120 Hz in spite of the same difference in the fundamental pitch frequency in both cases, and also by the fact that, in Fig. 2, P(c) for the pair of voice samples with fundamental pitch frequency of 160 Hz and 140 Hz is larger than P(c) for that of 140 Hz and 120 Hz.

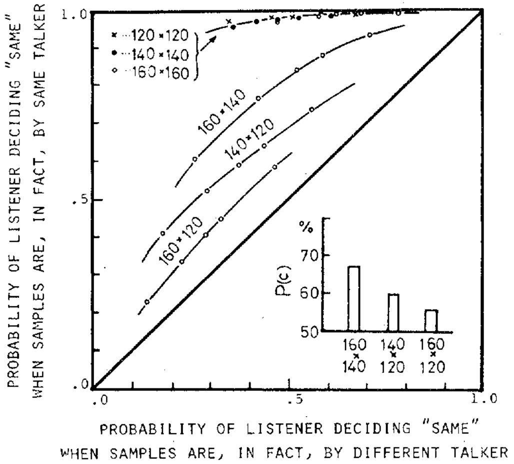

Fig. 2. Average ROC curves and P(c) of six listeners for every possible combination of three levels of mean fundamental pitch frequency in talker discrimination tests.

V. Relation between Acoustical Parameters and the Configuration on the PAS

As an approach to our ultimate goal of estimating the quantitative mapping function of the physical space onto the PAS, we first examined the relationship between the configuration on the PAS and the acoustical parameters in terms of a multiple correlation technique. The result of the multiple correlation calculation based on the three-dimensional configuration obtained in Section IV and the results of acoustical analysis in Section III for the 24 voice samples of voice set I is shown in Fig. 3. In this figure, the magnitude of the vector is linearly proportional to the multiple correlation coefficient, and its direction corresponds to the direction of maximum correlation.

From Fig. 3(b), it is clear that the mean fundamental pitch frequency is highly correlated with the configuration in the direction of the A_3 axis and, at the same time, is perceptually independent of other acoustical parameters that are mostly correlated with the configuration on the A_1-A_2 plane (beyond the 5 percent level of confidence). Among those parameters whose projection on the A_1-A_2 plane are shown in Fig. 3(a), the slope of glottal source spectrum is related to the positive direction of the A_1 axis, while the rapid fluctuation of pitch periods is nearly related to the negative A_1 axis. The formant frequencies, except for F_3, are related to A_2 axis as well as to A_1 axis, suggesting that the contribution of the vocal tract characteristics to the personal quality is somewhat different from that of the glottal source characteristics.

In order to examine the relative contribution of each of those acoustical parameters to perceptual cues

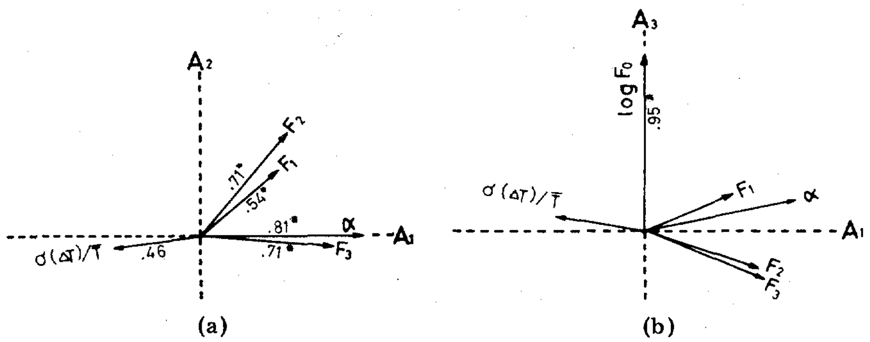

Fig. 3. Multiple correlation between the acoustical parameters and the configuration of voice set I, expressed in vector on A_1-A_2 and A_1-A_3 planes of the three-dimensional PAS. The numerals at the vectors indicate the multiple correlation coefficients and "*" indicates that the value is significant at 5 percent level of confidence. (a) Projections on the A_1-A_2 plane. (b) Projections on the A_1-A_3 plane.

TABLE I
Relation Between Various Sets of the Acoustical Parameters and the Explained Variance of the Configuration of Voice Set I on the PAS

GLOTTAL SOURCE CHARACTERISTICS			VOCAL TRACT CHARACTERISTICS	EXPLAINED VARIANCE IN %
MEAN FUNDAMENTAL PITCH FREQUENCY	FLUCTUATION OF FUNDAMENTAL PITCH PERIOD	SLOPE OF GLOTTAL SOURCE SPECTRUM	FORMANT FREQUENCIES, F_1, F_2 AND F_3	
X	X	X	X	86
X	X		X	84
X		X	X	84
X			X	81
X	X	X		71
X				55

of talker discrimination, it is necessary to take account of the differences in variance of the mapped points of the voice samples in each direction of maximum correlation to the acoustical parameters in addition to multiple correlation coefficient. So, the mapping function of the physical space onto the PAS was approximated by a linear multiple regression model, and the configuration of the voice samples on the PAS was estimated from various sets of acoustical parameters and explained variance (or the difference of total variance and residual variance normalized by the total variance) was calculated as shown in Table I.

As much as 55 percent of the total variance can be explained by the mean fundamental pitch frequency alone and 16 percent more can be explained by adding the slope of the glottal source spectrum and the fluctuation of the fundamental pitch period. If, on the other hand, the lowest three formant frequencies are added to the mean fundamental pitch frequency, 26 percent more can be explained than by the mean fundamental pitch frequency alone. Thus, it is shown that the mean fundamental pitch frequency plays an important role in the perception of personal quality, and that the relative contribution of the vocal tract characteristics to the personal quality is larger than that of the glottal source characteristics other than the mean fundamental pitch frequency. All together, 86 percent of the total variance is explained by those six acoustical parameters.

VI. Related Experiment Using Hybrid Voices

Some of those acoustical parameters, such as the third formant frequency and the slope of glottal source spectrum, whose multiple correlations are in a direction similar to the result of Section V, are correlated with each other in these voice samples. The correlation coefficient was 0.53, which is beyond a 1 percent level of confidence, and is considered to be much larger than that, due to the artifacts in the acoustical analysis. In this case, it is impossible to find with this limited set of natural voice samples how that pair of parameters is utilized in the perception of personal quality of voice.

In order to examine the differences in the contribution to the perception of personal quality between the slope of glottal source spectrum and the third formant frequency, an experiment similar to Miller's [7] was conducted with hybrid voice samples. In order to produce voice samples without natural correlation between those parameters, the source wave or the formant frequency pattern was interchanged in several ways among the five talkers' voice samples from voice set I that had a fundamental pitch frequency of exactly 140 Hz. This was done using a computer-programmed terminal analog speech synthesizer. The glottal source wave was generated by repeating the waveform of one pitch period that was derived from the inverse transform of the glottal source spectrum obtained in Section III for each talker. As a result, the fluctuation in fundamental pitch frequency characteristic of each talker was removed in the voice samples in this experiment.

An experiment similar to the one described in Section IV was conducted with another group of six listeners to obtain the dissimilarities among the voice samples. The configuration of the voice samples was derived from the dissimilarities between every possible combination of the 13 voice samples (${}_{13}C_2$ = 78), each of which was the average rate of "different talker" response for each of the two identical pairs (order reversed) for six listeners and ten trials ($2 \times 6 \times 10 = 120$ responses).

As it was shown in the previous experiment that the acoustical parameters other than the fundamental pitch frequency related primarily to the plane (A_1-A_2) in the three-dimensional configuration of voice set I and all of the voice samples in this experiment had the same fundamental pitch frequency, a two-dimensional configuration with 7.1 percent stress was selected to compare with the configuration of voice set I (Fig. 1). The axes of this two-dimensional configuration were rotated [8], dilated, and translated so as to match on a least squares basis the configuration of the five natural voice samples of /a/ with that obtained for the same five voice samples on the A_1-A_2 plane in Fig. 1.

The configuration of these synthetic voice samples on the A_1-A_2 plane of the PAS is shown in Fig. 4. In this result, the direction of the maximum correlation of the third formant frequency became similar to that of the first and second formant frequencies, while the direction of the maximum correlation of the slope of glottal source spectrum is almost diagonal to that of the third formant frequency. These trends are shown typically by the fact that the stimulus points of the voice samples consisting of the vocal tract characteristics of the original voice sample "6," which has the highest formant frequencies, are located at the most positive side of the direction of the maximum correlation of formant frequencies, while those consisting of the glottal source characteristics of the original voice sample "6," which has the gentler falling slope of glottal source spectrum, are located at the more positive side of the maximum correlation of the slope of glottal source spectrum.

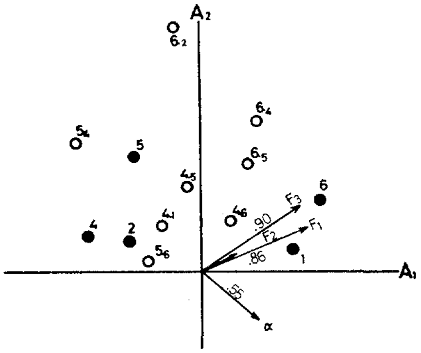

Fig. 4. Configuration on A_1-A_2 plane of the PAS for natural voice samples of vowel /a/ (closed circles) uttered by five male talkers (indicated by numerals) with fundamental pitch frequency of 140 Hz and synthetic voice samples (open circles) in which the formant frequency pattern of the talkers (indicated by larger numerals) and the glottal source characteristics of other talkers (indicated by smaller numerals) were combined.

Moreover, the finding in Section V that the relative contribution of the vocal tract characteristics is greater than that of the glottal source characteristics other than the mean fundamental pitch frequency was confirmed by the fact that the mapped points of the hybrid voice samples tended to be closer to those of the original voice samples having the same formant pattern. These results agree with the Miller's [7].

VII. Personal Quality in Different Vowels

To examine whether or not the results of the experiment with vowel /a/ are also found to hold for different vowels, an experiment similar to the one described in Section IV for the vowel /a/ was conducted with 40 stimuli consisting of the five Japanese vowels with a fundamental pitch frequency of 164 Hz uttered by each of eight male adult talkers (voice set II). Five of the eight talkers were common to both experiments. Every possible pair of the voice samples was presented 3 times to a group of 13 listeners, all of whom were different from those in the previous experiments. The configuration of voice set II was derived from the dissimilarities

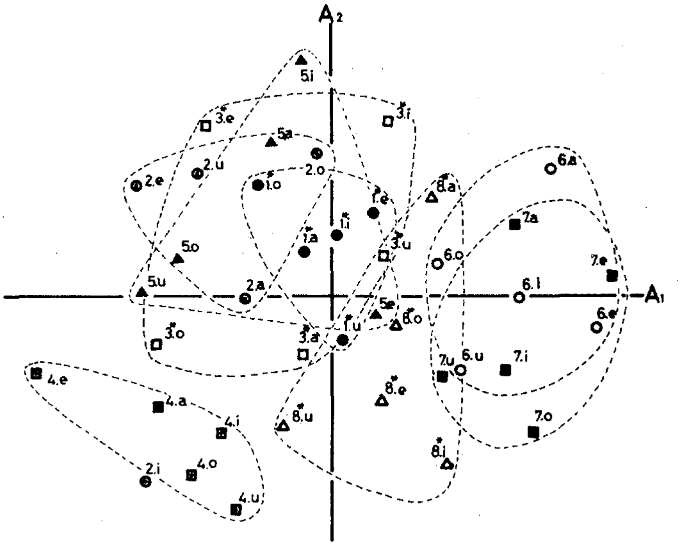

Fig. 5. Configuration on A_1-A_2 plane of the PAS for voice set II, which consists of the five Japanese vowels, /i,e,a,o,u/, uttered by eight male talkers (indicated by numerals and encircled by dotted lines) with fundamental pitch frequency of 164 Hz. Talkers 1*, 3*, and 8* are different from those of voice set I. Stress is 22.9 percent.

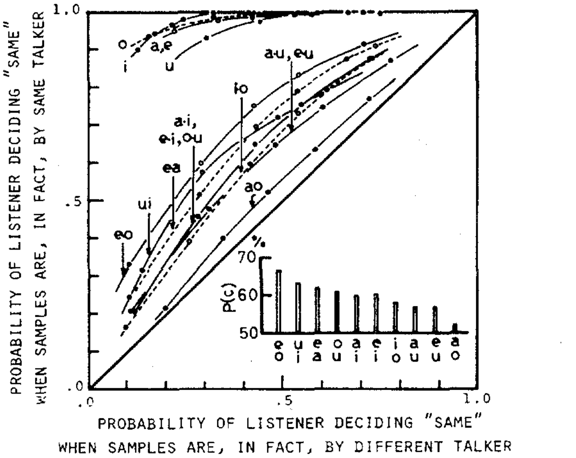

Fig. 6. Average ROC curves and P(c) of 13 listeners for every possible combination of the five Japanese vowels in talker discrimination tests.

between every possible combination of the 40 voice samples (${}_{40}C_2 = 780$), each of which is the average rate of "different talker" response for each of the 2 identical pairs (order reversed) for 13 listeners and 3 trials ($2 \times 13 \times 3 = 78$ responses).

For the same reason as in the previous section, the two-dimensional configuration of voice set II as shown in Fig. 5 was used to compare with the A_1-A_2 plane of the three-dimensional configuration of voice set I. The axes were rotated, dilated, and translated in the same way as described in Section VI. As the two-dimensional configuration of voice set II and that of voice set I on the A_1-A_2 plane correspond well with each other, this configuration was considered to represent reasonably the gross nature of the dissimilarities of personal quality among the voice samples of voice set II in spite of rather high stress (38.2, 22.9, 16.8, and 12.5 percent in one, two, three, and four dimensions, respectively). The ROC curves and the P(c) for each combination of vowels are shown in Fig. 6.

Since the mapped points of voice samples of a single talker cluster together in a particular region on the A_1-A_2 plane and the region for each talker covers a wide area on the plane, as shown in Fig. 5, it is clear that the perceptual cues of the personal quality common to different vowels is involved in the listener's judgement. This is supported by the fact that the percent correct score P(c) for the combination of the different vowels ranges from 57 to 67 percent, except for the combination of /a/ and /o/ as can be seen in Fig. 6.

The multiple correlation coefficients between the acoustical parameters and the configuration of eight stimuli for each of the five Japanese vowels of voice set II are shown in Fig. 7. For each of two of the parameters, namely the slope of glottal source spectrum [Fig. 7(d)] and the deviation of the first formant frequency from the average value for all talkers for each kind of vowel [Fig. 7(a)], the multiple correlation coefficients have similar directions and reasonably large values for most vowels (an exception is /u/ in the deviation of the first formant frequency). The similar but not so remarkable trends are found in the case of rapid fluctuation of fundamental pitch period [Fig. 7(e)]. It may therefore be concluded that those three parameters (in addition to the mean fundamental pitch frequency in the general case) relate to the perceptual cues for talker discrimination independent of most kinds of vowels. Other parameters, such as F_2 and F_3 did not show a direction common to all of the vowels, presumably because the inter-talker variance of the deviation of the first formant frequency (normalized by the average of all talkers) in the voice sample used here was fairly large (14 percent) compared with the averaged intra-vowel variance of the deviation of each talker (normalized by the average first formant frequency of all talkers) (8 percent), so that the ratio of the former to the latter was reasonably large (1.7). On the other hand, the second formant frequency had small inter-talker variance of deviation (7 percent) with small intra-vowel variance of deviation (5 percent). The third formant frequency had medium inter-talker variance of deviation (11 percent) with large intra-vowel variance of deviation (8 percent), so that the two ratios were not large enough (1.4 and 1.2 for the second and the third formant frequencies, respectively) for both formants to serve as the personal information independent of the kind of vowels.

The explained variance (as described in Section VI) was calculated for each vowel and is shown in Table

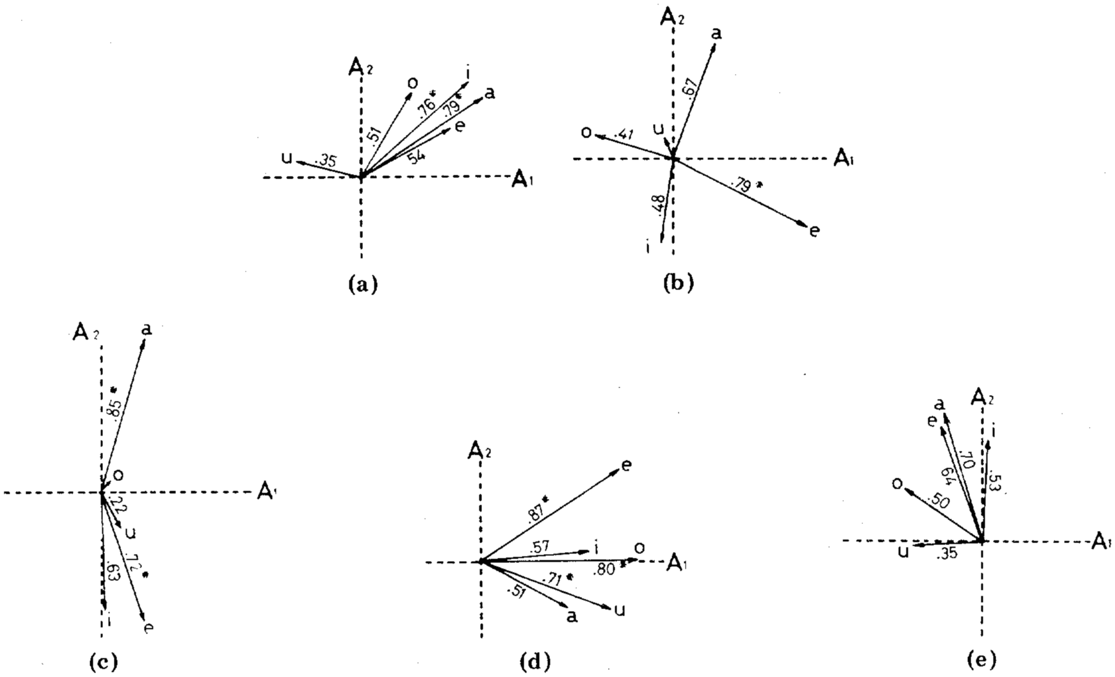

Fig. 7. Multiple correlation between the acoustical parameters and the configuration of voice set II for each of the five Japanese vowels, expressed in vector on A_1–A_2 plane of the PAS. The numerals at the vectors indicate the multiple correlation coefficients and "*" indicates that the value is significant at 5 percent level of confidence. (a) First formant frequency. (b) Second formant frequency. (c) Third formant frequency. (d) Slope of the glottal source spectrum. (e) Rapid fluctuation of pitch periods.

TABLE II
Relation Between Various Sets of the Acoustical Parameters and the Explained Variance of the Configuration of Voice Set II on the PAS

GLOTTAL SOURCE CHARACTERISTICS		VOCAL TRACT CHARACTERISTICS	EXPLAINED VARIANCE IN %				
FLUCTUATION OF FUNDAMENTAL PITCH PERIOD	SLOPE OF GLOTTAL SOURCE SPECTRUM	FORMANT FREQUENCIES, F_1, F_2 AND F_3	/i/	/e/	/a/	/o/	/u/
X	X	X	97	99	97	93	94
X	X		80	80	79	88	80
		X	71	80	85	62	45

II. 93 to 99 percent of total variance is explained when those five acoustical parameters are combined all together. With the two acoustical parameters of glottal source characteristics, the slope of glottal source spectrum and the rapid fluctuation of fundamental pitch frequency, about 80 percent of the configuration of voice set II can be explained whatever the vowel may be. Although the relative contribution of the vocal tract characteristics varies widely with each kind of vowel, it is smaller than that of the glottal source characteristics (other than mean fundamental pitch frequency) in most kinds of vowels and is larger in the case of vowel /a/. This is in agreement with the finding of the first experiment. The relative vocal tract contribution is largest in the case of /a/ (85 percent) and is very small in the case of /u/ and /o/ (45 percent and 62 percent, respectively). This is presumably because the contribution of the third formant frequency is less in the vowel /u/ and /o/ compared with other vowels as the energy of the third formant is smaller for those vowels having lower first and second formant frequencies.

VIII. Conclusions

In order to investigate quantitatively the relation between the perceptual difference in personal quality and the difference in the acoustical properties of sustained vowels, the dissimilarities among the voice samples based on the confusion in talker discrimination tests were represented as distances on the PAS. Then, the relation between the configuration of the voice samples on the PAS and their acoustical properties were examined by means of multiple correlation and regression analyses.

The results obtained from this study are summarized as follows.

1) The relative contribution of the mean fundamental pitch frequency to the perception of the personal quality of voice is the largest among all parameters, and its contribution to the perceptual dimension is almost independent of those of other acoustical parameters. (Although those results were obtained from the experiment with the vowel /a/, they may be extended to the case of other vowels since the fundamental pitch frequency is almost independent of formant frequencies in the mechanism of speech production and perception in general.)

2) Among the voice samples with same fundamental pitch frequency, the vocal tract characteristics (the deviation of the formant frequencies) and the glottal source characteristics (the slope of the glottal source spectrum and the rapid fluctuation of the fundamental pitch period) contribute to different perceptual dimensions from each other. The magnitude of the contribution of the vocal tract characteristics to the perceptual difference of personal quality varies widely according to the kind of vowel, while this is not the case with the glottal source characteristics. Only in the case of the vowel /a/ is the contribution of the vocal tract characteristics larger than that of the glottal source characteristics.

3) Among those acoustical parameters whose perceptual dimensions are independent of most kinds of vowel are the deviation of the first formant frequency from the typical value for each kind of vowel, the slope of the glottal source spectrum, and the rapid fluctuation of the fundamental pitch periods, in which, of course, included in general is the mean fundamental pitch frequency.

Although the results of this experiment depend on to what extent the listeners were actually judging personal quality and to what extent just difference in sound quality, we think that the PAS obtained from this experiment shows the multidimensional nature of the perception of personal quality that was intended to be investigated, based on following reasons. The points that represent the voice samples uttered with significantly different pitch frequencies, or the voice samples of different kinds of vowels, cluster together in small regions characteristic of each talker on the A_1-A_2 plane of the PAS. Furthermore, talker discrimination was carried out correctly with a P(c) of about 60 to 70 percent for the two voice samples in a pair whose difference in sound quality caused by the difference in fundamental pitch frequency or in the kind of vowels was significantly larger than the just noticeable difference. These considerations indicate that the listeners were judging the voice samples not by the perceptual difference of sound quality but by that of the personal quality of voice.

The rather high stress obtained for the configurations in these experiments is due to the error in the dissimilarity data caused by averaging the rate of "different talker" response for all listeners, rather than to the lack of number of dimensions. In addition, the stress tends to increase with an increase in the number of dissimilarity data in the general case. So these configurations cannot be regarded as "unlikely to be of interest" by judging only from the amount of stress.

On the other hand, the configurations of the voice samples of the five talkers common to the three experiments, but utilizing different listener groups, correspond to each other fairly well. This fact will support the idea that most of the perceptual difference involved in the dissimilarity data was represented in the three-dimensional PAS (the two-dimensional PAS for the voice samples of the same fundamental pitch frequency).

Considering the fact that the greatest part of the perceptual difference of personal quality was explained by the five acoustical parameters used in this study, and that the vowels contribute most of the energy in speech, it may be said that those acoustical parameters play important roles in the perception of the static nature of the personal quality in speech.

It will be necessary, in the future, to make clear the effects of some of the other parameters, such as the zero pattern in the glottal source spectrum and forment bandwidth, in addition to above-mentioned parameters on the static nature of the personal quality, as well as to investigate the dynamic nature of the personal quality.

Acknowledgment

The authors wish to thank Prof. Kido for his valuable cooperation and assistance to this study.

References

[1] H. Kasuya, Y. Kanamori, S. Arai, and K. Kido, "Psychological auditory space representing vowel quality," in *Proc. 7th Int. Cong. Acoust.*, Budapest, 1971, Paper 20C5.
[2] M. V. Mathews, J. E. Miller, and E. E. David, Jr., "Pitch

synchronous analysis of voiced sounds," *J. Acoust. Soc. Amer.*, vol. 33, pp. 179-186, Feb. 1961.

[3] D. W. Marquardt, "Algorithm for least square estimation of nonlinear parameters," *J. Soc. Ind. Appl. Math.*, vol. 11, p. 431, 1963.

[4] A. J. Compton, "Effects of filtering and vocal duration upon the identification of speakers, aurally," *J. Acoust. Soc. Amer.*, vol. 35, p. 1748, 1963.

[5] J. B. Kruskal, "Nonmetric multidimensional scaling: A numerical method,"*Psychometrica*, vol. 29, p. 115, 1964.

[6] J. P. Egan, A. I. Schuman, and G. Z. Greenberg, "Operating characteristics determined by binary decisions and by ratings," *J. Acoust. Soc. Amer.*, vol. 31, p. 768, 1959.

[7] J. E. Miller, "Decapitation and recapitation, a study of voice quality," *J. Acoust. Soc. Amer.*, vol. 36, p. 2002 (A), 1965.

[8] N. Cliff, "Orthogonal rotation to congruence," *Psychometrica*, vol. 31, pp. 33-42, 1966.

Helium Speech Unscramblers—A Critical Review of the State of the Art

THOMAS A. GIORDANO, HOWARD B. ROTHMAN,
and HARRY HOLLIEN

Abstract—The development of saturation diving has enabled man to work in the sea at great depths and for long periods of time. This advance has resulted, in part, as a consequence of the substitution of helium for nitrogen in breathing gas mixtures. However, the utilization of HeO_2 breathing mixtures at high ambient pressures has caused problems in speech communication; in turn, electronic aids have been developed to improve diver communication. These helium speech unscramblers attempt to process variously the grossly unintelligible speech resulting from the effects of helium-oxygen breathing mixtures and ambient pressure, and to reconstruct such signals in order to provide adequate voice communication. This paper presents a discussion of the effects of HeO_2/P on speech and then describes some of the techniques used to "unscramble" the distorted speech. Included among the techniques are: 1) frequency subtraction; 2) tape recorder playback; 3) vocoder approaches; 4) digital coding; and 5) convolution processing. In addition, a generalized evaluation of these approaches is included.

Introduction

The development of saturation diving is permitting man to work in the sea for long periods of time—and at great depths. This situation, coupled with the need for more working divers of all types, has resulted in the explosive expansion of the numbers of individuals engaged in such activity. However, divers experience difficulty working cooperatively due to the lack of adequate diver-to-diver and diver-to-surface communication. Consequently, there is currently a critical need for the development of good communication techniques and related equipment; such a thrust is a vital one if divers are to operate at their full potential.

The Communication Sciences Laboratory at the University of Florida is involved in a program of basic and applied research designed to investigate some of the problems faced by man working in the sea. It is common knowledge [2], [3], [7], [8], [11], [14], [16], [19], [26]-[31] that speech communication is severely affected the moment the diver submerges beneath the surface. It has been established further that, with increasing depth (a situation commonly found in saturation diving), unique problems in speech communication arise due to the complex effects of the diver breathing helium/oxygen (HeO_2) gas mixtures at high ambient pressures. The use of HeO_2 breathing mixtures is necessary in order to avoid the narcotic and, ultimately, toxic effects of nitrogen at great depths.

As stated, the investigators cited above have established that a diver talking in the HeO_2/P environment experiences severe speech distortion. Electronic aids (usually referred to as HeO_2 unscramblers) have been developed to cope with this problem. However, these units are not yet capable of processing the distorted speech to levels considered adequate for good voice communication [13], [24]. Further, the theoretical constructs upon which unscrambler design is based have not been adequately developed. Therefore, in order to evaluate HeO_2 speech unscramblers and as one phase of our research program in diver communication, we have undertaken a four-part project designed to: 1) determine the exact nature of this type of equipment (such data *often* is not available due to proprietary rights); 2) develop standardized tests for evaluating all types of HeO_2 speech unscramblers; 3) evaluate unscramblers on-line; and 4) evaluate them off-line. The present paper deals primarily with the first thrust of the four listed; essentially it constitutes a review of the nature of HeO_2 speech unscramblers currently in use. In order to do so in a meaningful

Manuscript received January 2, 1972; revised April 4, 1973. This work was supported by the Physiological-Psychology and Engineering Psychology Branches of the Office of Naval Research.

T. A. Giordano is with the Army Electronics Command, Fort Monmouth, N. J.

H. B. Rothman and H. Hollien are with the Communication Sciences Laboratory, University of Florida, Gainesville, Fla. 32601.