論文 / 著書情報
Article / Book Information

論題(和文)	
Title	Perception of Voice Individuality and Physical Correlates
著者(和文)	古井 貞熙
Author	Sadaoki Furui, Masahito Akagi
出典(和文)	日本音響学会 聴覚研究会資料, Vol. , No. , pp. H-85-18
Journal/Book name	, Vol. , No. , pp. H-85-18
発行日 / Issue date	1985,

音声における個人性の知覚と物理関連量

PERCEPTION OF VOICE INDIVIDUALITY AND PHYSICAL CORRELATES

古井 貞熙　赤木 正人

Sadaoki FURUI　Masato AKAGI

日本電信電話公社武蔵野電気通信研究所

Musashino Electrical Communication Laboratory, N.T.T.

内容梗概：音声に含まれる個人性情報の知覚に寄与する物理的特性として、時間平均スペクトル包絡と時間平均ピッチ周波数を取り上げ、心理実験にもとづく検討を行った。男性話者９名の単語音声を受聴したときの話者間の心理的距離と、物理的特徴による距離との関係を２種類の単語について調べた結果、種々のスペクトル包絡の抽出法のうち、ケプストラムで平滑化したスペクトル包絡が最も心理量との対応が良く、平均ピッチ周波数とともに、心理的距離との高い相関が得られた。さらに、スペクトル包絡の個人差は主として２．５～３．５kHzの帯域の平均レベルにあり、心理量と良い対応を示すことがわかった。

I. INTRODUCTION

Speech waves convey linguistic (phonetic) information, speaker-dependent (personal) information, and many other kinds of information. Among these, personal information plays the most important role next to the linguistic one. Personal information takes the form of voice quality, voice height, loudness, speed, tempo, intonation, accent, vocabulary, and so on. Various kinds of physical features having complicated interactions produce these voice characteristics. Voice quality and height, which are the most important of the auditory types of personal information, can be related mainly to the spectral envelope and fundamental frequency (pitch).

Itoh et al.[1] evaluated the contribution of the spectral envelope, pitch, and dynamic characteristics of these features on the perception of voice individuality, using short sentences and five sustained vowels which were processed in various ways by a PARCOR analysis-synthesis system. They found that the spectral envelope is most influential, the pitch is second, and the dynamic characteristics are third. Interactions were also found between the spectral envelope and pitch and between the spectral envelope and the dynamic characteristics.

Matsumoto et al.[2] calculated the auditory inter-speaker similarity matrix based on the stimulus-stimulus confusion rates in the speaker descrimination test. In such a test, vowels spoken by different speakers were presented in pairs. They applied multi-dimensional scaling (MDS) to the similarity matrix, and investigated the relationship between this solution and three kinds of physical features, that is, (1) from the first to the third formant frequencies (F1-F3), (2) inclination of the estimated vocal-cord spectrum, and (3) the mean value and fluctuation of the pitch frequency. Among these features, all of which are relevant to voice-individuality, they concluded that the mean pitch frequency has the largest effect.

Kuwabara et al.[3] also calculated the psychological inter-speaker distances based on the results of the seven-step psychological similarity rating for the paired word stimuli uttered by the different speakers. In this experiment, the word, /aoiue/, which consists of only vowels, was used. They examined the correlation coefficients between the psychological inter-speaker distances and the physical distances obtained by (1) the time function of pitch, (2) the mean pitch frequency, (3) the time function of F1 through F5, and (4) the spectral envelope at the vowel centers. Their results indicate that the pitch-related features of (1) and (2) exhibit relatively large correlation coefficients, whereas the spectrum-related features of (3) and (4) exhibit small values.

Kuwabara et al.[4] have also analyzed the acoustic features of the speech of several announcers and found that their speech can be characterized by the dynamic characteristics of pitch and formant time functions. They also found that the voices of the announcers have a higher spectral level at the frequency band between 3 and 4 kHz, compared to people of other occupations. They assumed that the high spectral level gives a penetrating quality to the voices of announcers in the same way as that the "singing formant" does to the voices of singers.

As stated above, most results of research on the relationship between psychological voice individuality and physical characteristics indicate that the pitch has a larger correlation with the psychological characteristics than the spectral envelope. However, the spectral envelope is thought to be making the same or an even greater contribution to the perceptual voice individuality. In automatic speaker recognition, the techniques based on the spectral envelope are usually more effective than those based on the pitch[6,7].

With the idea that the importance of the spectral envelope has been underestimated due to inappropriate envelope estimation techniques in the previous research, we have evaluated several kinds of spectral representation (estimation) techniques from the viewpoint of correlation with perceptual voice individuality. The results have also been compared with the pitch information. Although the dynamic features of speech also convey large amounts of voice individuality[8,9], only the text-independent time-averaged (static) features are investigated in this paper, since the time-averaged information has been regarded as conveying the basic individuality[10,11].

II. PSYCHOLOGICAL EXPERIMENTS

2.1. ACQUISITION OF DISSIMILARITY DATA

A psychological experiment was performed using word utterances by male speakers to obtain the psychological inter-speaker distance (dissimilarity) data which indicate the psychological differences of voice quality. The rating method, a kind of direct judgment method was used in the experiment. In this method, word utterances by two speakers were sequentially presented and listeners were asked to evaluate the similarity of the two utterances by using one of the five categories: "similar", "slightly similar", "difficult to decide", "slightly dissimilar", and "dissimilar".

There are several other methods for acquiring the psychological distances, of which we will discuss three. The first is a paired comparison (method of triads or method of tetrads), in which a pair having a smaller or larger dissimilarity is required to be selected. The second is a ranking method, in which each stimulus in turn represents a standard, while the remaining stimuli are to be placed in the order of rank with respect to their degree of similarity to the standard. Finally, is a method based on the confusion rate when listeners are required to identify the speaker of an utterance, or to choose whether the speakers of two utterances are the same or not.

There are several reasons for using the rating method in this paper. First, since natural speech samples were used in the experiment, there were a considerable number of pairs which were too different to be confused. Second, it is desirable that the number of pairs presented to listeners be reduced as much as possible to decrease the burden on the listeners. Third, since the stimuli can be presented only sequentially in the hearing experiment, it is undesirable for a set of utterances presented to the listeners to have so many stimuli that the utterances to be compared are separated in time. Fourth, it is desirable that a considerable amount of data be obtained in a short time and that the data be processed efficiently.

2.2. STIMULI AND METHOD OF EXPERIMENTS

Two kinds of Japanese word utterances, /namae/ and /baNgo:/, which were uttered naturally by nine male speakers in an unechoic room, were used in the experiment. A single utterance by each speaker was recorded for each word. The nine speakers in their mid-twenties to mid-fifties, were selected arbitrarily. The utterances, passed through a low-pass filter having a 3.4 kHz cut-off frequency, were initially stored in a computer storage, and then presented to the listeners.

Seventy-two pairs of stimuli for each word were presented four times for each listener. There were 16 listeners, consisting of 12 male and four female listeners. The stimuli were presented to one ear through a receiver (DR-305) in an unechoic room at a comfortable speech level (roughly 0 dB-OTR). The listeners were instructed to evaluate the voice quality while ignoring the difference in loudness and accent. The inter-stimulus interval (ISI) and the inter-test-interval (ITI) were 0.3 sec and 2 sec, respectively.

2.3. EXPERIMENTAL RESULTS

The dissimilarity matrix for the voice quality of the nine male speakers was obtained by averaging the rating scores for all listeners for each pair of stimuli. The rating scale was assigned as follows; 0:"similar", 1:"slightly similar", 2:"difficult to decide", 3:"slightly dissimilar", and 4:"dissimilar".

The relationship between the elements of dissimilarity matrices for /namae/ and /baNgo:/ is presented in Fig.1. The correlation coefficient for this relation is 0.76. The correlation coefficient between the rating scores for /namae/ and /baNgo:/ is distributed among the listeners between 0.44 and 0.97, whose mean value is 0.61.

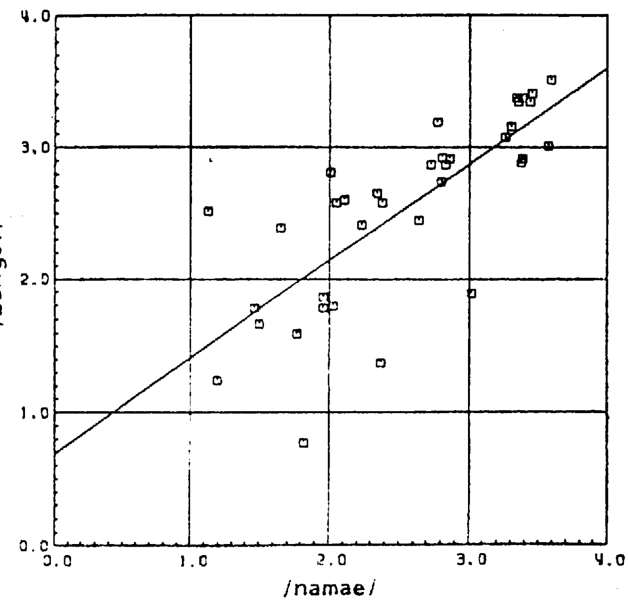

Fig. 1. Relationship between the psychological inter-speaker distances for word /namae/ and /baNgo:/.

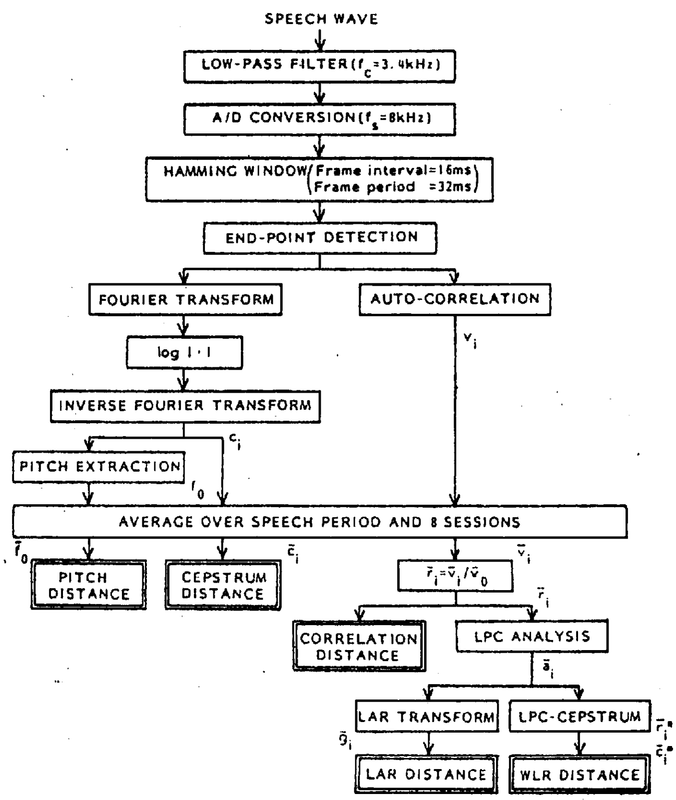

Fig. 2. Block diagram indicating the principal operations for distance calculations.

III. CONSTRUCTION OF PHYSICAL DISTANCE MEASURE

3.1. AVERAGED SPECTRAL ENVELOPE AND PITCH FREQUENCY EXTRACTION

The spectral envelope and pitch frequency averaged over the speech period of each word were extracted as physical correlates having perceptual voice individuality. The relationship between the inter-speaker distance calculated using the physical features and the psychological distance obtained through the hearing experiment was investigated. Since it has already been made clear that the physical features of individual voice fluctuate over a long duration such as several months[10-13], those averaged over eight sessions of three-month intervals (averaged over a period of nearly two years) were used in this paper.

The principal operations for the physical distance calculations are indicated in Fig.2. The speech wave is passed through a low-pass filter having a 3.4 kHz cut-off frequency in the same way as in the psychological experiment. It is then processed by A/D conversion at the sampling frequency of 8 kHz. A 32 msec frame period is extracted every 16 msec by applying the Hamming window. Additionally, the beginning and end of the actual utterance are determined through the mean power for each frame. In the left path of Fig.2, Fourier, logarithmic and inverse Fourier transformations are applied to each frame of the detected actual utterance period to extract the 0th through the 11th cepstrum coefficients (c_i). Pitch frequency (f_0) is extracted using higher order cepstrum coefficients. In the right path of Fig.2, the 0th through the 12th order auto-correlation functions (v_i) are calculated.

Cepstrum and auto-correlation coefficients and pitch frequency are then averaged over the eight sessions after being averaged over the speech period of each utterance. Each speaker is then represented by the physical features, $\bar{c}_i$, $\bar{v}_i$ and $\bar{f}_0$. The inter-speaker distance matrix is calculated using either of these features.

The averaged auto-correlation functions are normalized by the 0th order function (power), and converted into auto-correlation coefficients ($\bar{r}_i$).

Linear predictive coding (LPC) analysis is applied to the auto-correlation coefficients. Inter-speaker distances using the log area ratios (LARs, $\bar{g}_i$) and the weighted likelihood ratio (WLR) distance using the LPC-cepstrum ($\bar{c}_i^*$) and modified correlation coefficients ($\bar{r}_i^*$) are based on the LPC analysis technique. The LPC-cepstrum, that is, the cepstrum obtained through LPC analysis (LPC model), as well as correlation coefficients used in the WLR distance calculation are extracted after spectral equalization. This equalization is performed by the first-order inverse filtering (fixed to all speakers) using the first-order correlation coefficient averaged over all of the speakers.

3.2. INTER-SPEAKER DISTANCE MEASURES

The five inter-speaker distance measures evaluated in this paper are defined as

(1) cepstrum distance, $D_1(j,k)=\sqrt{\sum_{i=1}^{p1}(\bar{c}_{ij}-\bar{c}_{ik})^2}$,

(2) correlation coefficient distance, $D_2(j,k)=\sqrt{\sum_{i=1}^{p2}(\bar{r}_{ij}-\bar{r}_{ik})^2}$,

(3) log area ratio (LAR) distance, $D_3(j,k)=\sqrt{\sum_{i=1}^{p3}(\bar{g}_{ij}-\bar{g}_{ik})^2}$,

(4) weighted likelihood ratio (WLR) distance, $D_4(j,k)$

$$=\sqrt{\sum_{i=1}^{p4}(\bar{r}_{ij}^*-\bar{r}_{ik}^*)\,(\bar{c}_{ij}^*-\bar{c}_{ik}^*)} ,$$

(5) and pitch frequency distance, $D_5(j,k)=|\bar{f}_{0j}-\bar{f}_{0k}|$,

where, j and k are the indices of the speakers. The number of parameters, p1-p4, were set to 11 or 12 based on the previous results of research on speech analysis-synthesis systems. In these systems, synthesized speech which maintains voice individuality can be realized using not more than 10 correlation coefficients when the speech is bandlimited under 3.4 kHz.

IV. RELATIONSHIP BETWEEN PSYCHOLOGICAL AND PHYSICAL DISTANCES

4.1. CORRELATION WITH SPECTRAL ENVELOPE DISTANCES

Correlation coefficients between the inter-speaker distance matrices which are calculated by the distance measures defined in the previous chapter and the psychological distance matrix (dissimilarity) are presented in Table 1. This table shows that the cepstrum distance exhibits the maximum correlation with the psychological distance, whose correlation coefficient is 0.70 for a two-word average. Since 95% confidential intervals of the correlation coefficient are between 0.63 and 0.83 for /namae/ and between 0.50 and 0.77 for /baNgo:/, the cepstrum and psychological distances are considered to be correlative. The relationship between the cepstrum and psychological distances for 72 pairs of speakers is presented in Fig.3. The WLR and LAR distances exhibit smaller correlations with the psychological distance, and the correlation coefficient distance has a much smaller correlation value as well.

Table 1-Correlation coefficients between psychological and physical distances

Distance	Word /namae/	/baNgo:/
Cepstrum (D1)	0.74	0.66
Correlation (D2)	0.47	0.27
LAR (D3)	0.57	0.37
WLR (D4)	0.58	0.42

4.2. CORRELATION WITH PITCH FREQUENCY DISTANCE

The relationship between the inter-speaker distance using pitch frequency and the psychological distance is presented in Fig.4. The correlation coefficients are 0.71 for /namae/ and 0.76 for /baNgo:/, which are nearly equal to the correlation coefficient between the cepstrum and psychological distances.

Correlation coefficients between the two physical distances, that is, the pitch frequency and cepstrum distances, are 0.38 for /namae/ and 0.59 for /baNgo:/. This indicates that the two physical distances exhibit some correlation in the word /baNgo:/, whereas no correlation can be observed in the word /namae/. It can therefore be said that pitch frequency is as important as the spectral envelope in representing the perceptual voice individuality.

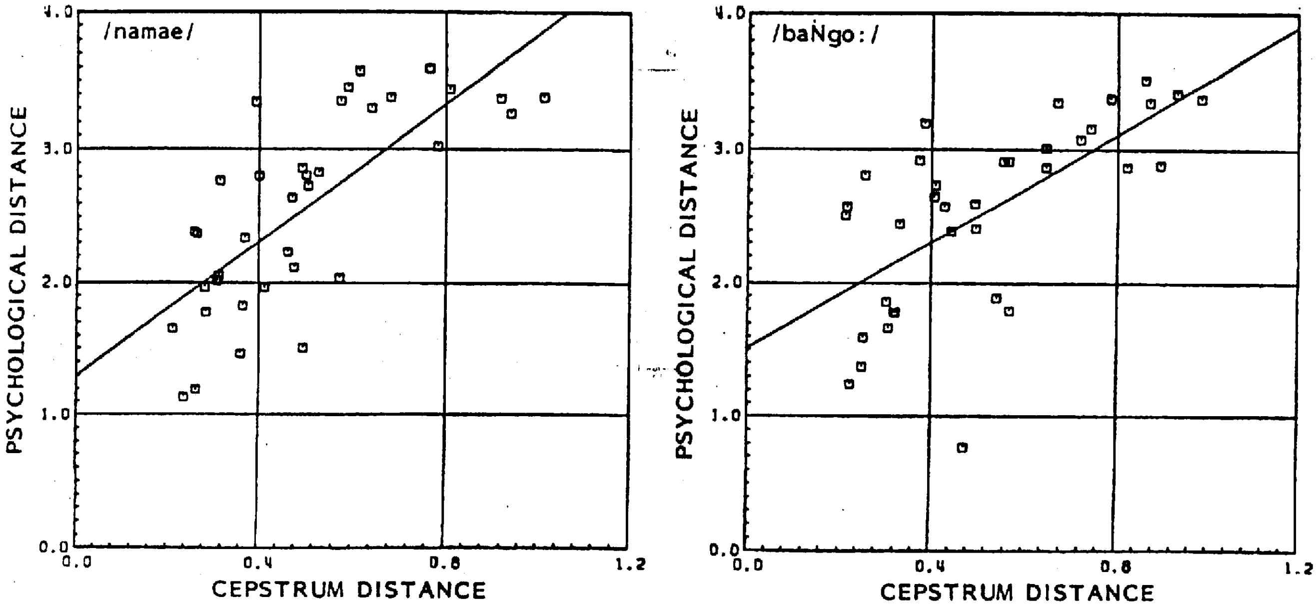

Fig. 3. Relationship between the psychological distance (dissimilarity) and cepstrum distance.

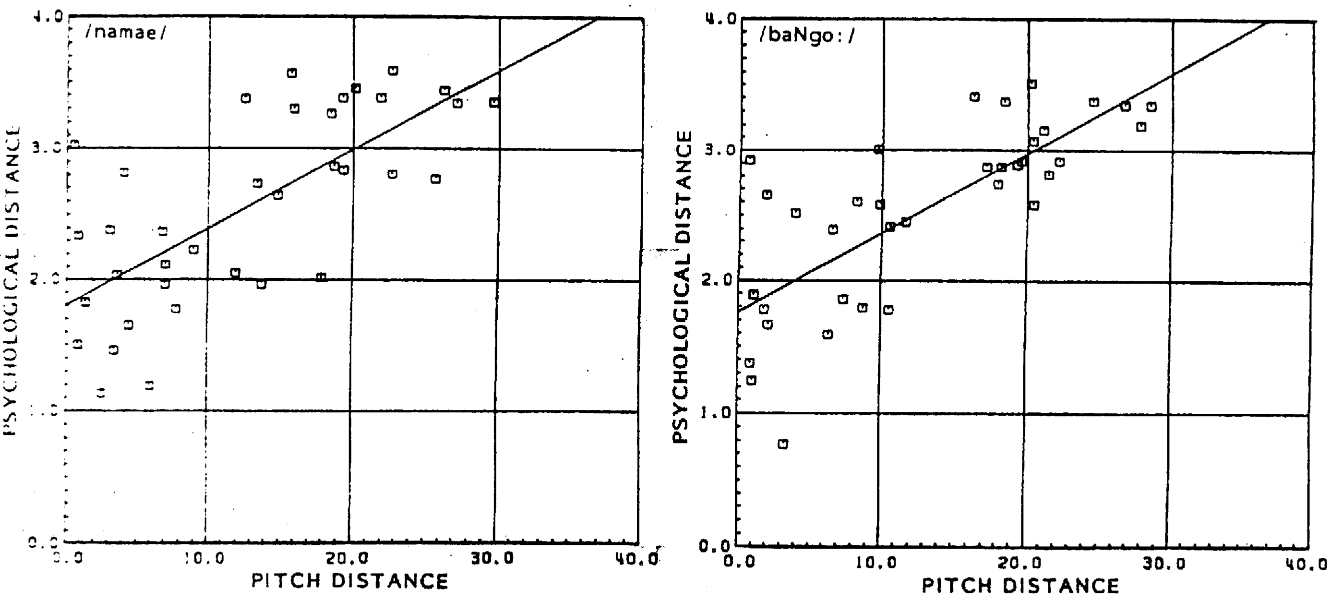

Fig. 4. Relationship between the psychological distance and pitch frequency distance.

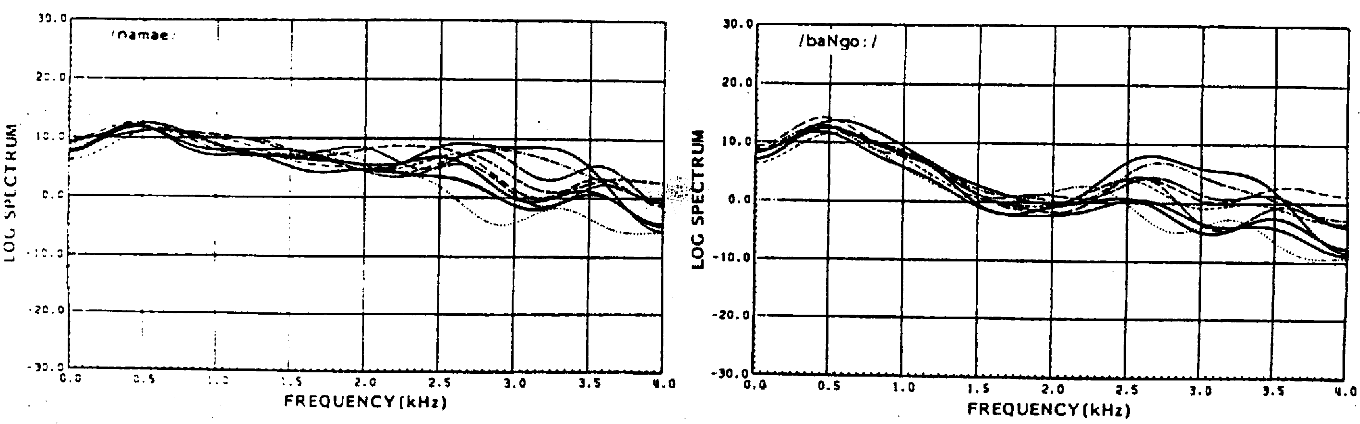

Fig. 5. Spectral envelopes derived from the averaged cepstrum coefficients for the voices of nine speakers.

4.3. SPECTRAL ENVELOPE REPRESENTED BY THE CEPSTRUM

Since the cepstrum is defined by the inverse Fourier transformation of the logarithmic spectrum, a smoothed logarithmic spectral envelope can be obtained by the Fourier transformation of the lower order 10 cepstrum coefficients or so. Hence, the cepstrum distance corresponds to the distance of the smoothed spectral envelopes. The spectral envelopes for the nine speakers derived from the cepstrum are shown in Fig.5. This figure indicates that the individuality of the spectral envelopes derived from the cepstrum mainly exists in the difference of the spectral levels of the frequency band roughly between 2.5 and 3.5 kHz.

Based on this result, the mean level of the spectral envelope between these values was calculated for each speaker as the main "spectral feature" of voice individuality. On the other hand, each speaker was projected onto a two-dimensional plane which was obtained by the MDS (Kruskal's non-metric method) applied to the cepstrum distance matrix. Multiple-regression analysis was then attempted between the projection and the "spectral feature".

The projection of each speaker is shown in Fig.6, and the "spectrum" in this figure indicates the axis which maximizes the correlation between the "spectral feature" and the regression of each speaker. The correlation coefficients are 0.99 for both /namae/ and /baNgo:/. Kruskal's stresses for these two-dimensional planes are 2.8 % and 0.8 % for /namae/ and /baNgo:/ respectively. It can be concluded that the individuality of the spectral envelope derived from the cepstrum is almost completely represented by the difference of spectral levels between 2.5 and 3.5 kHz.

The MDS is then applied to the psychological distance matrix to obtain a two-dimensional plane on which the nine speakers were projected. The multiple-regression analysis is applied between the projection and the two kinds of physical features, the pitch frequency and the "spectral feature". The results are presented in Fig.7, which indicate the large correlation (0.91-0.95) between the physical and psychological features. Kruskal's stresses for these planes are 6.6 % and 8.3 % for /namae/ and /baNgo:/ respectively. This figure also implies a correlation between the "spectral feature" and the pitch frequency in /baNgo:/. Several other physical features remain to be investigated to explain the component on the plane of Fig.7 that can not be explained by either the "spectral feature" or the pitch frequency.

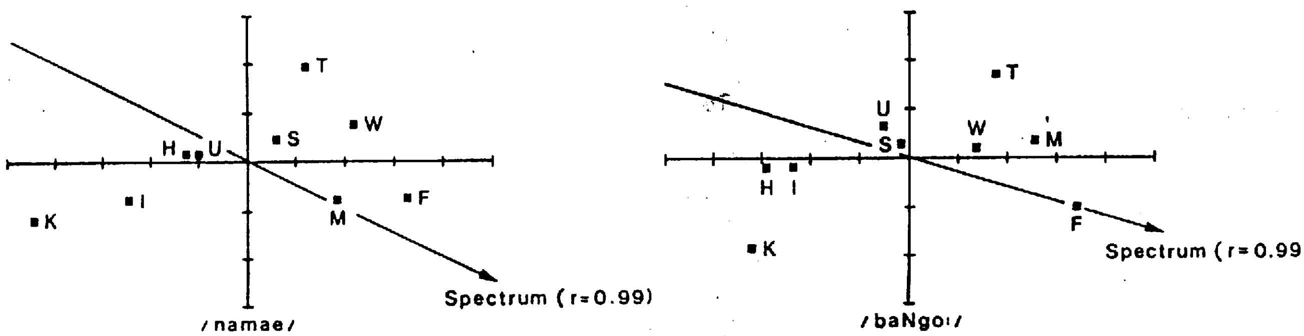

Fig. 6. Projections of the nine speakers onto the plane produced by MDS applied to the inter-speaker cepstrum distance matrix. "spectrum" indicates the direction which maximizes the correlation between the regression of each speaker to the vector and the mean spectral level of the 2.5 to 3.5 kHz frequency band.

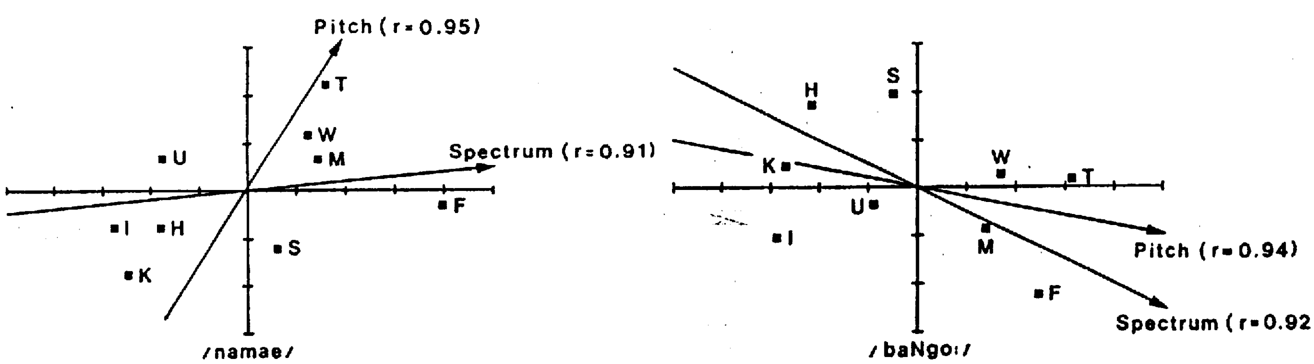

Fig. 7. Projection of the nine speakers onto the MDS plane for psychological distances. "spectrum" and "pitch" indicate the directions which maximize the correlation between the regression of each speaker to the vector and the physical properties. Spectrum: mean spectral level of frequency band between 2.5 and 3.5 kHz; pitch: pitch frequency.

4.4. COMPARISON WITH LPC SPECTRAL ENVELOPE

The cepstrum distance and other spectral envelope distances evaluated in this paper have the two differences indicated in Table 2. In order to separate the effects of the two items, a supplementary experiment was attempted using distances other than the cepstrum distance. In this experiment, the correlation functions were averaged after power normalization. However, the modification of the averaging procedure caused only a slight difference and produced no increase in the correlation between the physical and psychological distances. This means that the difference between the cepstrum and other distances can be attributed to the characteristics of spectral representation.

Table 2-Differences between cepstrum distance and other distances

	Cepstrum distance	Other distances
Parameter averaging	Averaging after power normalization	Power-weighted averaging
Spectral represen-tation	Equally weighted to peaks and dips, and smoothed	Sensitive to dips (corr.coef.) or peaks (LAR & WLR)

The spectral envelopes derived through the LPC analysis which are the basis of the LAR and WLR distances were then plotted for each speaker as shown in Fig.8. In comparison with Fig.5, Fig.8 indicates that the spectral peaks in the low frequency region, which are mainly related to the phonetic information, still remain in the LPC spectral envelopes following the time averaging process. The spectral peaks obscure the individuality of the spectral envelopes. On the other hand, only the global spectral features remain in the spectral envelopes derived from the time-averaged cepstrum, in which formant effects are reduced by smoothing.

The above results are consistent with the automatic speaker recognition results[6,9,14] in which smoothed spectral envelopes derived from the cepstrum are much more effective than those obtained through LPC analysis.

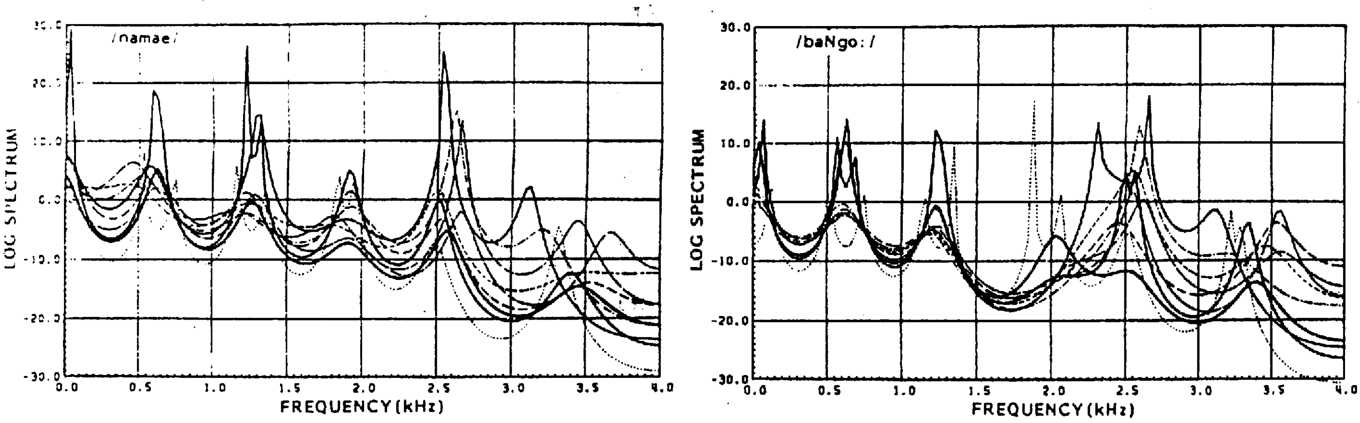

Fig. 8. Spectral envelopes derived through the LPC analysis for the voices of nine speakers.

V. SUPPLEMENTARY EXPERIMENT BY COMPLETE METHOD OF TRIADS[15]

A supplementary experiment was performed based on Torgerson's complete method of triads using the same word /namae/ uttered by the nine male speakers. The purpose was to confirm the reliability of the results of the above psychological experiments. In the complete method of triads, three stimuli are successively presented to each listener with some inter-stimulus intervals. The listeners are then required to select either one of the second or third stimulus which is relatively similar to the first. The relative distances between the stimuli are calculated based on these responses. Since $n(n-1)(n-2)/6$ triads are generally made for n kinds of stimuli, 252 kinds of triads were made for the nine speakers voices. The triad were presented to eight male speakers, two of whom were the same listeners used in the previous experiment.

The correlation coefficient between the psychological inter-speaker distance (dissimilarity) obtained with this method and that with the previous rating method was 0.79. The relationship to the cepstrum distance is given in Fig.9, whose correlation coefficient is 0.69. This result confirms that the psychological distances which this paper is based on have high reliability.

VI. DISCUSSION AND CONCLUSION

This paper has revealed that the spectral envelope obtained through time-averaging and frequency-smoothing using cepstrum coefficients, as well as averaged pitch frequency, has a close connection with the psychological individuality of voice quality. It has also been made clear that inter-speaker distances using the spectral envelope have almost the same correlation with the psychological distances as those using the pitch frequency. Individuality of the spectral envelope mainly exists in the mean level of the frequency band between 2.5 and 3.5 kHz. This corresponds highly with the psychological characteristics. This result is similar to the analysis results for the voices of announcers and singers, which means that the spectral level around 3 kHz is relevant to a wide range of general voice individuality, and not restricted to special professional voices.

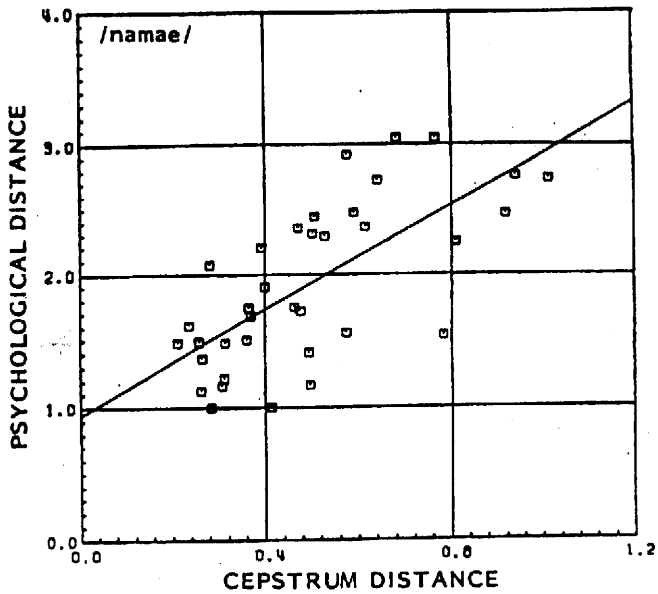

Fig. 9. Relationship between the psychological distance obtained by the complete method of triads and the cepstrum distance.

Previous statistical analysis of the individuality included in the cepstrum of longtime-averaged spectra extracted from short sentences[10] indicated that relatively large amounts of individual information are included in the first, third and fourth cepstrum coefficients, when the sampling frequency is 8kHz. The first-order cepstrum coefficient corresponds to the inclination of the spectral envelope, whereas the third- and fourth-order coefficients correspond to the global spectral pattern which repeats with a 2 to 2.7 kHz period. The analysis results for the longtime-averaged spectra thus highly correspond with the psychological results reported in this paper.

As stated in the Introduction, many kinds of physical features appear to be related to voice individuality. Although it has been made clear that the time-averaged features of the spectral envelope and pitch frequency are basically important, it is also clear that these features can not explain all of the aspects of individuality. Further investigation, current and projected, includes an analysis of dynamic spectral features and a detailed and large-scale investigation based on a wide range of utterances.

ACKNOWLEDGMENT

The authors wish to thank Kazuhiko Kakehi, Head of the Communication Principles Research Section 4, and other members of the section for their stimulating discussions. We also wish to acknowledge the cooperation given by Kazuo Ueda, Faculty of Literature, Kyoto University, and the listeners.

REFERENCES

[1] K.Itoh and S.Saito:"Effects of acoustical feature parameters of speech on perceptual identification of speaker",Trans.IECE,J65-A,1, pp.101-108 (1982)
[2] H.Matsumoto,et al.:"Multidimensional representation of personal quality of vowels and its acoustical correlates",IEEE Trans.,AU-21,5, pp.428-436 (1973)
[3] H.Kuwabara and K.Ohgushi:"Psychological similarity of natural speech and its physical correlates",Fall ASJ Meeting 1-2-9 (1981)
[4] H.Kuwabara and K.Ohgushi:"Acoustic characteristics of professional male announcers' speech", Trans.IECE,J66-A,6,pp.545-552 (1983)
[5] J.E.Sundberg:"Articulatory interpretation of the 'singing formant'",J.Acoust.Soc.Amer.,55,4, pp.838-844 (1974)
[6] B.S.Atal:"Automatic recognition of speakers from their voices",Proc.IEEE,64,4,pp.460-475 (1976)
[7] S.Furui:"Review and perspective of speaker recognition",Journ.IECEJ,67,5,pp.537-543 (1984)
[8] S.Saito and S.Furui:"Personal information in dynamic characteristics of speech spectra",Proc. 4th IJCPR,pp.1014-1018 (1978)
[9] S.Furui:"Cepstral analysis technique for automatic speaker verification",IEEE Trans,ASSP-29,2,pp.254-272 (1981)
[10] S.Furui,et al.:"Talker recognition by longtime averaged speech spectrum",Trans.IECE,55-A, 10,pp.549-556 (1972)
[11]S.Furui:"Comparison of speaker recognition methods using statistical features and dynamic features",IEEE Trans,ASSP-29,3,pp.342-350 (1981)
[12]S.Furui and F.Itakura:"Talker recognition by statistical features of speech sounds",Trans. IECE,56-A,11,pp.717-724 (1973)
[13]S.Furui:"An analysis of long-term variation of feature parameters of speech and its application to talker recognition",Trans.IECE,57-A,12, pp.880-887 (1974)
[14]S.Nakagawa and M.Sakamoto:"Evaluation of FFT cepstrum and LPC cepstrum for speech and speaker recognition",Trans.IECE,J66-A,12,pp.1199-1206 (1983)
[15]S.Furui:"Speaker-dependent features of speech signal",Mathematical Sciences,252, pp.33-41 (1984)