# RESEARCH ON INDIVIDUALITY FEATURES IN SPEECH WAVES AND AUTOMATIC SPEAKER RECOGNITION TECHNIQUES

Sadaoki FURUI

*Musashino Electrical Communication Laboratory, Nippon Telegraph and Telephone Corporation, Musashino-shi, Tokyo, 180 Japan*

Received 23 November 1985

**Abstract.** This paper presents an overview of Japanese research on individuality information in speech waves, which have been performed from various points of view. Whereas physical correlates having perceptual voice individuality have been investigated from the psychological viewpoint, research from the engineering viewpoint is related to automatic speaker recognition, speaker-independent speech recognition, and training algorithms in speech recognition.

Speaker recognition research can be classified into two classes, depending on whether or not the text is predetermined. However, it has been made clear that even if the text is not predetermined, text-dependent individual information can be used that is based on explicit or implicit phoneme recognition. Various examples of speaker recognition methods are classified into these variations, and their performances are presented in this paper. In particular, this paper focuses on the long-term intra-speaker variability of feature parameters as on of the most crucial problems in speaker recognition.

Additionally, it presents an investigation into methods for reducing the effects of long-term spectral variability on recognition accuracy.

**Zusammenfassung.** Der folgende Artikel bietet einen Überblick über die Arbeiten in Japan welche sich mit der sprecherspezifischen Information im Sprachsignal befassen. Die Wechselbeziehung zwischen physikalishen Merkmalen und der wahrgenommenen Individualität der Stimme is vom psychologischen Standpunkt aus untersucht worden. Wohingegen, die technischen Wissenschaften sich mehr für die automatische Sprechererkennung, für die automatische sprecherunabhängige Spracherkennung, sowie für das automatische Präparieren von Spracherkennungssystemen anhand von Prototypen interessiert haben.

Die Arbeiten über die Sprechererkennung als solche können in zwei Klassen eingeteilt werden, abhängig davon ob der Text fixiert ist oder nicht. Es ist nachgewiesen worden, anhand von ausdrücklicher oder impliziter Phonemerkennung, dass auch wenn der Text nicht vorherbestimmt ist, textabhängige Sprecherinformation nutzbar ist. Verschiedene Beispiele von Sprechererkennungsmethoden werden in diesem Beitrag vorgestellt. Im besonderen wird das Problem der sprecherspezifischen Laugzeitvariabilität der akustischen Merkmale hervorgehoben. Dieses Problem muss als eines der kritischsten im Rahmen der Sprechererkennung bezeichnet werden.

Dieser Beitrag stellt eine Untersuchung an über Methoden welche es erlauben die Auswirkungen der langzeitvariabilität spektraler Grössen auf die Erkennungsgenauigkeit zu reduzieren.

**Résumé.** Cet article présente un aperçu des recherches entreprises au Japon, relatives à l'information individuelle véhiculée par l'onde de parole. Alors que les corrélats physiques des traits perceptifs de l'identité de la voix ont été étudiés du point de vue psychologique, la recherche menée du point de vue de l'ingénieur se rattache à la reconnaissance automatique du locuteur, à la reconnaissance de la parole indépendante du locuteur, et aux algorithmes d'apprentissage en reconnaissance de la parole.

La recherche en reconnaissance du locuteur peut être catalogueé en deux classes, selon que le texte est prédéterminé ou non. Cependant, il a été mis en évidence que, même si le texte n'est pas prédéterminé, l'information individuelle dépendante du texte peut être utilisée et ce, sur base de la reconnaissance explicite ou implicite du phonème. Divers exemples de méthodes de reconnaissance du locuteur sont classées ici selon ces variantes et leurs performances sont présentées. En particulier, cet article éclaire la variabilité à long terme des paramètres intra-locuteur comme l'un des problèmes les plus cruciaux en reconnaissance du locuteur.

En plus, cet article présente une étude des méthodes destinées à réduire les effets de la variabilité spectrale à long terme sur la précision de la reconnaissance.

**Keywords.** Individuality, speaker recognition, text-dependent, text-independent, intra-speaker variability, spectrum.

## 1. Introduction

Speech waves convey linguistic (phonetic) information, speaker-dependent (individual) information, and many other kinds of information. Among these, individual information plays the most important role next to linguistic information. Individual information takes the form of voice quality, voice height, loudness, speed, tempo, intonation, accent, the use of 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 individual information, can be related mainly to the static and dynamic characteristics of the spectral envelope and fundamental frequency (pitch).

Speaker recognition is the process of automatically deciding on the speaker, and is based on several physical features representing the individual information extracted from the speech wave. Speaker recognition research, which is closely related to the techniques concerning speaker-independent speech recognition, is currently being performed from different perspectives. Speaker recognition can be divided principally into speaker verification and speaker identification. Speaker verification is the process of accepting or rejecting the identity claim of a speaker by comparing a set of measurements of the speaker's utterances with a reference set of measurements of the utterance of the person whose identity is claimed. Speaker identification is the process of determining which of the registered speakers a given utterance comes from.

Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former require the speaker to issue a predetermined utterance, whereas the latter do not rely on the specific text being spoken.

When speaker verification systems are realized, voice can be used as a key to verify the identity claim of a speaker in an extensive array of services such as banking transactions using a telephone network, database acquisition services including personal information, and security control for confidential information areas. Several systems are currently planned for future information-age application.

## 2. Individuality information in the speech spectrum

The following linear structure model, which has a phoneme factor, a speaker factor, and a phoneme × speaker interaction factor, is assumed to represent the $p$-dimenstional vector of a spectrum at time $t$, that is,

$$X(t) = (x_1(t), x_2(t),\ldots,x_p(t)):$$

$$X_{ijk}(t) = \boldsymbol{\mu}(t) + \boldsymbol{\alpha}_i(t) + \boldsymbol{\beta}_j(t) + \gamma_{ij}(t) + \boldsymbol{\varepsilon}_{ijk}(t), \quad (1)$$

where $k = 1,\ldots,r$,

$$\sum_{i=1}^{a} \boldsymbol{\alpha}_i(t) = \sum_{j=1}^{b} \boldsymbol{\beta}_j(t) = \sum_{i=1}^{a} \gamma_{ij}(t) = \sum_{j=1}^{b} \gamma_{ij}(t) = 0, \quad (2)$$

and

$$\boldsymbol{\varepsilon}_{ijk} \sim N(\mathbf{0},\boldsymbol{\Lambda}), \quad (3)$$

where $\boldsymbol{\alpha}_i(t)$ is the $i$th main effect of the phoneme factor, $\boldsymbol{\beta}_j(t)$ is the $j$th main effect of the speaker factor, $\gamma_{ij}(t)$ is the $(i,j)$th effect of the interaction of both factors, and $r$ is the number of observation iterations. $\boldsymbol{\mu}(t)$ is decided so that Eq. (2) can be satisfied. $\boldsymbol{\varepsilon}_{ijk}(t)$ is assumed to be distributed independently under the $p$-dimensional normal distribution. $a$ and $b$ are the number of levels for the phoneme factor ($i = 1,\ldots,a$) and the speaker factor ($j = 1,\ldots,b$), respectively. When the above-mentioned assumptions are satisfied, the effect of each factor can be tested on the basis of the multivariate analysis of variance using $\chi^2$ distributions [1]. The parameter $t$ is omitted hereafter for simplicity.

Using isolated five-vowel utterances recorded through a 600-type telephone connected to an artificial subscriber line or through a high-quality microphone, the author [2,3] has analyzed the effect of each factor based on the above-mentioned structure model. Each vowel was uttered by 18 male speakers three times during each of two sessions separated by an interval of one year (six samples were used for each vowel for each speaker). In order to normalize the differences of the number of levels for each factor, each one was evaluated by its logarithmic likelihood ratio divided by the critical value for a certain confidence level. These normalized likelihood ratios were seen as being related to a wide-ranging recognition accuracy.

The speech waves were passed through a low-pass filter with a 3.4 kHz cut-off frequency, and sampled at 8 kHz. The time functions of the first- to 12th-order cepstrum coefficients extracted from the digitized speech were averaged across the entire speech interval for each utterance. The likelihood ratios were separately analyzed on the basis of the averaged cepstrum values for telephone speech, microphone speech, and mixed speech.

Each of these values divided by the critical value at the significance level of 1% is indicated in Fig. 1. As can be seen, the phoneme effect is overwhelmingly large compared to the speaker effect, which is second in size. It should also be noted that the interaction is also fairly large. All the effects for telephone speech are smaller than those for microphone speech, and all of these, especially the speaker effect and interaction, become smaller still when microphone and telephone speech are analyzed together. These results indicate that the extraction of the speaker factor is more difficult than that of the phoneme factor.

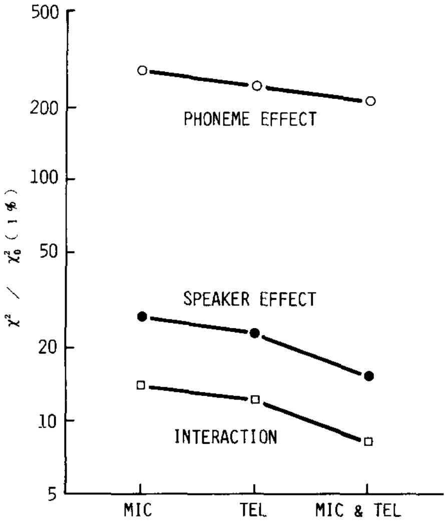


Fig. 1. Effects of phoneme, speaker, and interaction factors divided by critical values at the significance level of 1% in vowel utterances by 18 male speakers. Utterances were simultaneously recorded through a hig-quality microphone and a telephone set and were analyzed separately or jointly.

The correlation between the individual information in different vowels was examined using isolated five-vowel utterances by 15 male and 15 female speakers as a different approach to the individual information commonly included in different vowel spectra. Each speaker uttered each vowel only once. Linear predictive coding (LPC) cepstrum coefficients were extracted for the spectrum averaged over the interval around each vowel center, and were Fourier-transformed into a log-spectral envelope. A 30×30 inter-speaker distance matrix consisting of the distances between the spectral envelopes at the frequency band between 0 and 5 kHz for each pair of speakers was calculated for each vowel. The correlation coefficients between distance matrices for each pair of vowels were then calculated. The results, the averaged value of which is 0.49, are shown in Table 1. It can be concluded that the individual information in different vowels is correlative, since the reliable region for the confidence level of 95% is between 0.42 and 0.57 when the calculated correlation coefficient is 0.49.

When eq. (1)–(3) are satisfied, text-independent individual information can be extracted by

$$\mathbf{X}_{\cdot j \cdot} = E\left\{\frac{1}{a}\sum_{i=1}^{a}\mathbf{X}_{ijk}\right\} = \boldsymbol{\mu} + \boldsymbol{\beta}_j. \tag{4}$$

This corresponds to the calculation of the long-time averaged speech spectrum, the process of which is shown in Fig. 2(a). This method, which will be described in detail in Subsection 4.1, is disadvantageous in that it requires averaging over a long period to extract a stable $\boldsymbol{\beta}_j$. This is because the speaker effect is smaller than the phoneme

Table 1
Correlation coefficients between inter-speaker distance matrices for each pair of vowels

| | /a/ | /i/ | /u/ | /e/ | /o/ |
|---|---|---|---|---|---|
| /a/ | | 0.48 | 0.40 | 0.48 | 0.49 |
| /i/ | | | 0.59 | 0.66 | 0.52 |
| /u/ | | | | 0.45 | 0.45 |
| /e/ | | | | | 0.42 |
| /o/ | | | | | |

(average = 0.49)

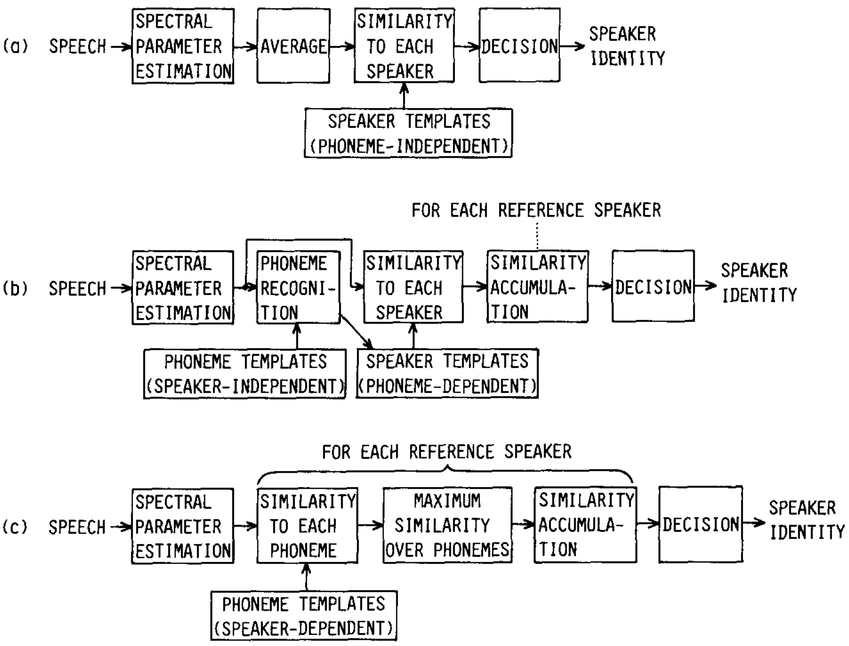

(a)
SPEECH
SPECTRAL PARAMETER ESTIMATION
AVERAGE
SIMILARITY TO EACH SPEAKER
DECISION
SPEAKER IDENTITY
SPEAKER TEMPLATES (PHONEME-INDEPENDENT)
FOR EACH REFERENCE SPEAKER
(b)
SPEECH
SPECTRAL PARAMETER ESTIMATION
PHONEME RECOGNI-TION
SIMILARITY TO EACH SPEAKER
SIMILARITY ACCUMULA-TION
DECISION
SPEAKER IDENTITY
PHONEME TEMPLATES (SPEAKER-INDEPENDENT)
SPEAKER TEMPLATES (PHONEME-DEPENDENT)
FOR EACH REFERENCE SPEAKER
(c)
SPEECH
SPECTRAL PARAMETER ESTIMATION
SIMILARITY TO EACH PHONEME
MAXIMUM SIMILARITY OVER PHONEMES
SIMILARITY ACCUMULA-TION
DECISION
SPEAKER IDENTITY
PHONEME TEMPLATES (SPEAKER-DEPENDENT)

Fig. 2. Three kinds of text-independent speaker recognition methods: (a) the method using the long-time averaged spectrum; (b) the method including explicit decision of phonemes, and (c) the method including implicit decision of phonemes.

effect and is in fact comparable to the interaction effect.

In order to cope with this disadvantage, two other methods for extracting individual information have also been examined. One involves the predetermination of a speech text or, in other words, level $i$ of the phoneme factor. The other concerns the extraction of the speaker factor following the automatic recognition of phoneme $i$. The former method corresponds to-text-dependent speaker recognition. The latter method can be divided into two variations depending on whether $i$ is explicitly (see Fig. 2(b)) or implicitly (see Fig. 2 (c)) recognized.

In the first variation, phoneme $i$ is recognized for every short segment of the speech wave on the basis of its similarity to speaker-independent phoneme reference templates. Its similarity to the reference template of each speaker associated with the recognized phoneme, i.e.,

$$\mathbf{X}_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \tag{5}$$

is then calculated. Similarity values for all short segments are averaged across the entire speech interval to produce an overall similarity to each reference speaker.

In the second variation, in which phoneme $i$ is not explicitly recognized, every short segment of input speech is compared with all the reference phoneme templates of all the speakers. Based on similarity values, a phoneme template exhibiting maximum similarity is selected for each reference speaker. These maximum similarity values are averaged over the entire speech interval for each reference speaker. The speaker with the maximum averaged value is then selected as the speaker of the input speech. These two variations will be described in detail in Subsections 4.2 and 4.3 respectively.

Although recognition methods similar to those already used in word recognition systems can be used in text-dependent speaker recognition, several feature extraction or weighting techniques are necessary to reduce as much as possible the variation introduced by the factors other than individuality. This reduction in variation enables the achievement of higher speaker recognition accuracy.

## 3. Physical characteristics related to the auditory perception of individuality

Matsumoto et al. [4] calculated the auditory inter-speaker similarity matrix based on the stimulus-stimulus confusion rates in a speaker-discrimination test. In the test, vowels spoken by different speakers were presented in pairs. Matsumoto et al. then applied multi-dimensional scaling (MDS) to the similarity matrix, and investigated the relationship between this solution and three kinds of physical features;

(1) from the first to the third formant frequencies ($F_1$–$F_3$);

(2) inclination of the estimated vocal-cord spectrum; and

(3) the mean value and fluctuation of the pitch frequency.

Of these features, all of which are relevant to voice individuality, they concluded that the mean pitch frequency has the largest effect.

Kuwabara et al. [5] 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 different speakers. Their experiment utilized the word, /aoiue/, which only consists of vowels. 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 funcion of $F_1$–$F_5$; 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. [6] also analyzed the acoustic features of the speech of several announcers, and found that their speech could be characterized by the dynamic characteristics of pitch and formant time functions. They also found that, compared to people of other occupations, announcers' voices had a higher spectral level at the 3–4 kHz frequency band. They assumed that the high spectral level gave a penetrating quality to the announcers' voices in the same way as the "singing formant" [7] does to the voices of singers.

Using utterances consisting of a short sentence and a connection of five sustained vowels, which were originally uttered by five male speakers and processed in various ways by a PARCOR analysis-synthesis system, Itoh et al. [8] evaluated the contribution of the spectral envelope, pitch, and dynamic characteristics of these features on the perception of voice individuality. Their experimental setup is shown in Fig. 3. Naturally spoken and synchronously spoken sentences were used in the experiments. The latter were spoken at the same time as a speaker gave a fixed guiding utterance which was presented to each speaker through a receiver in order to remove the individual information associated with dynamic characteristics, that is, the nonlinear expansion and contraction of sentence utterances.

The stimuli used in the experiments were:

(a) original speech;

(b) normal PARCOR analysis-synthesis speech;

(c) analysis-synthesis speech with fixed pitch frequency;

(d) analysis-synthesis speech using only source information (pitch, amplitude, and a voiced/unvoiced factor), where PARCOR coefficients were fixed to remove the spectral envelope information; and

(e) analysis-synthesis speech using only source amplitude information, where pitch and PARCOR coefficients were both fixed.

Twelve subjects were requested to identify the speaker of the input speech from five reference speakers. Figure 4 indicates the experimental results. The analysis of variance based on the identification rates for naturally and synchronously spoken sentences confirms that the main effect is largest for the spectral envelope, becomes smaller

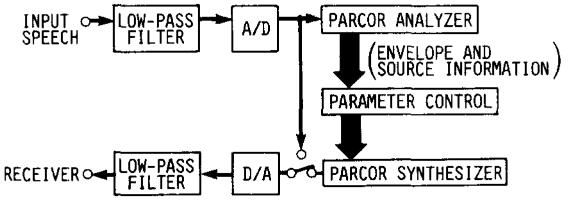
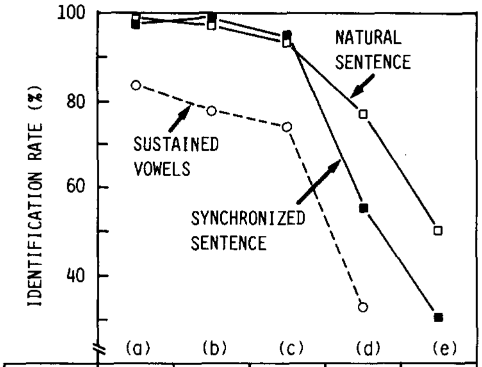


Fig. 3. Block diagram of the experimental system to evaluate the physical correlates of perceptual voice individuality using the PARCOR analysis-synthesis system.



| | | | | | |
|---|---|---|---|---|---|
| SPECTRUM | O | O | O | | |
| PITCH | O | O | | O | |

Fig. 4. Speaker identification rates for various experimental conditions: (a) original speech; (b) normal PARCOR analysis-synthesis speech; (c) analysis-synthesis speech with fixed pitch frequency; (d) analysis-synthesis speech using only source information, and (e) analysis-synthesis speech using only source amplitude information.

for the pitch, and smaller still for the dynamic characteristics. The interaction is significant both between the spectral envelope and the pitch, and that between the spectral envelope and the dynamic characteristics. These results mean that when the spectral envelope is maintained (conditions (a), (b) and (c) in Fig. 4), pitch and dynamic characteristics do not affect identification. Furthermore, when the spectral envelope is removed (conditions (d) and (e)), the identification score is largely decreased and is definitely affected by pitch and dynamic characteristics.

An additional experiment using voices synthesized by the combination of the spectral envelope and the driving source characteristics extracted from different speakers has revealed that the former characteristic plays a more important role than the latter.

By means of the relational analysis between these features and the results of a psychological experiment, the author et al. [9] investigated the time-averaged spectral envelope and pitch frequency as physical correlates having text-independent perceptual voice individuality. In the psychological experiment, inter-speaker distances between nine male speakers were obtained through a rating method using five categories located between "similar" and "dissimilar". Each speaker uttered two kinds of Japanese words, /namae/ and /baNgo:/, and these utterances were presented to 16 male and female listeners.

The physical parameters for each speaker were obtained by averaging the parameters over eight sessions at three-monthly intervals (averaged over a period of nearly two years). This was done since the physical features of voice have been observed to fluctuate over a long duration such as one of several months (see Subsection 4.1)

Experimental results for the relationship analysis between physical and psychological inter-speaker distances indicate that the distance for the spectral envelope smoothed by the first- to 12th-order cepstrum coefficients correlates more with the psychological distance than any other spectral envelope distance. As shown in Table 2, the averaged pitch frequency distance has almost the same correlation coefficient value with respect to the psychological distance as the spectral envelope has. The reliable regions for the 95% confidence level shown in the table indicate that these physical feature distances correlate with psychological distance. On the other hand, the two physical distances do not correlate well with each other.

A precise investigation into individuality in spectral envelopes has made it clear that the individuality mainly exists in the differences of the mean levels of the spectral envelopes between 2.5 and 3.5 kHz, as shown in Fig. 5, and that it is highly correlated with the psychological value. 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

Table 2
Correlation coefficients between psychological and physical inter-speaker distances
( ): Reliable region for confidence level of 95%

| Physical | Word | |
|---|---|---|
| features | /namae/ | /baNgo:/ |
| Spectral envelope | 0.74<br>(0.63–0.83) | 0.66<br>(0.50–0.77) |
| Pitch frequency | 0.71<br>(0.57–0.80) | 0.76<br>(0.64–0.84) |

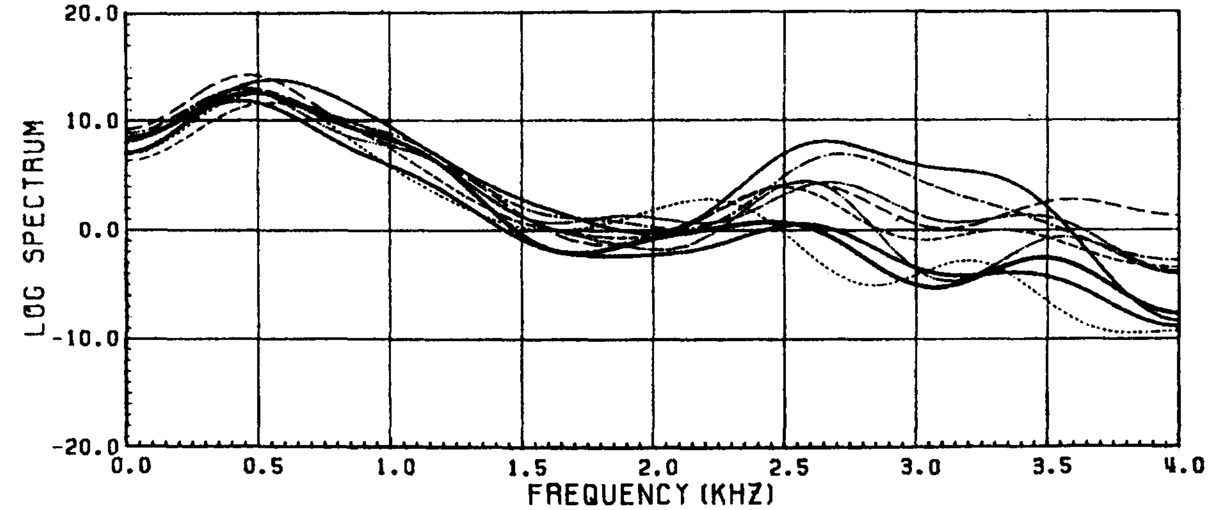


Fig. 5. Spectral envelopes derived from the averaged cepstrum coefficients for the word /baNgo:/ uttered by nine male speakers.

a wide range of general voice individuality, and not simply restricted to special professional voices.

## 4. Text-independent speaker recognition

### 4.1. Speaker recognition by long-time averaged spectrum

One of the typical examples of the method shown in Fig. 2(a) is that based on the long-time averaged spectrum (LAS). The author et al. [10] have expanded the LAS by cepstrum coefficients in order to analyze and normalize its long-term intra-speaker variations. Short spoken sentences of roughly 10 s in length were sampled at 8 kHz and an LAS was calculated by averaging the short-time spectrum extracted every 32 ms throughout the entire speech interval. The LAS was then transformed into cepstrum coefficients following energy normalization. Lower order cepstrum elements are commonly known to be related to the global pattern of the logarithmic spectrum, whereas higher order ones are related to the microscopic structure of the spectrum. This means that both spectral features can be processed separately. Additionally, this method is advantageous in that cepstrum distance is proportional to the distance of logarithmic spectral envelopes smoothed by the cepstrum.

Figure 6 compares the variance of each cepstrum coefficient for the LAS calculated for long-term speech data for a 12 month period, and for short-term data for a 10 day period. This

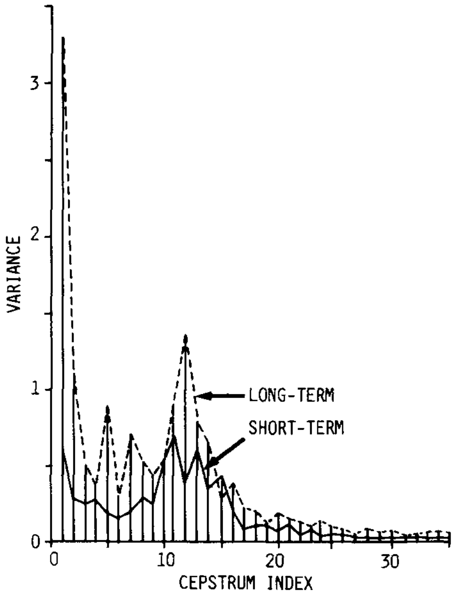
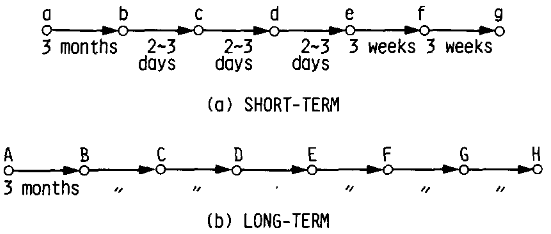


Fig. 6. Variance of each cepstrum coefficient for long-time averaged spectrum, calculated for long-term speech data of 12 months period and short-term data of 10 days period.



(a) SHORT-TERM

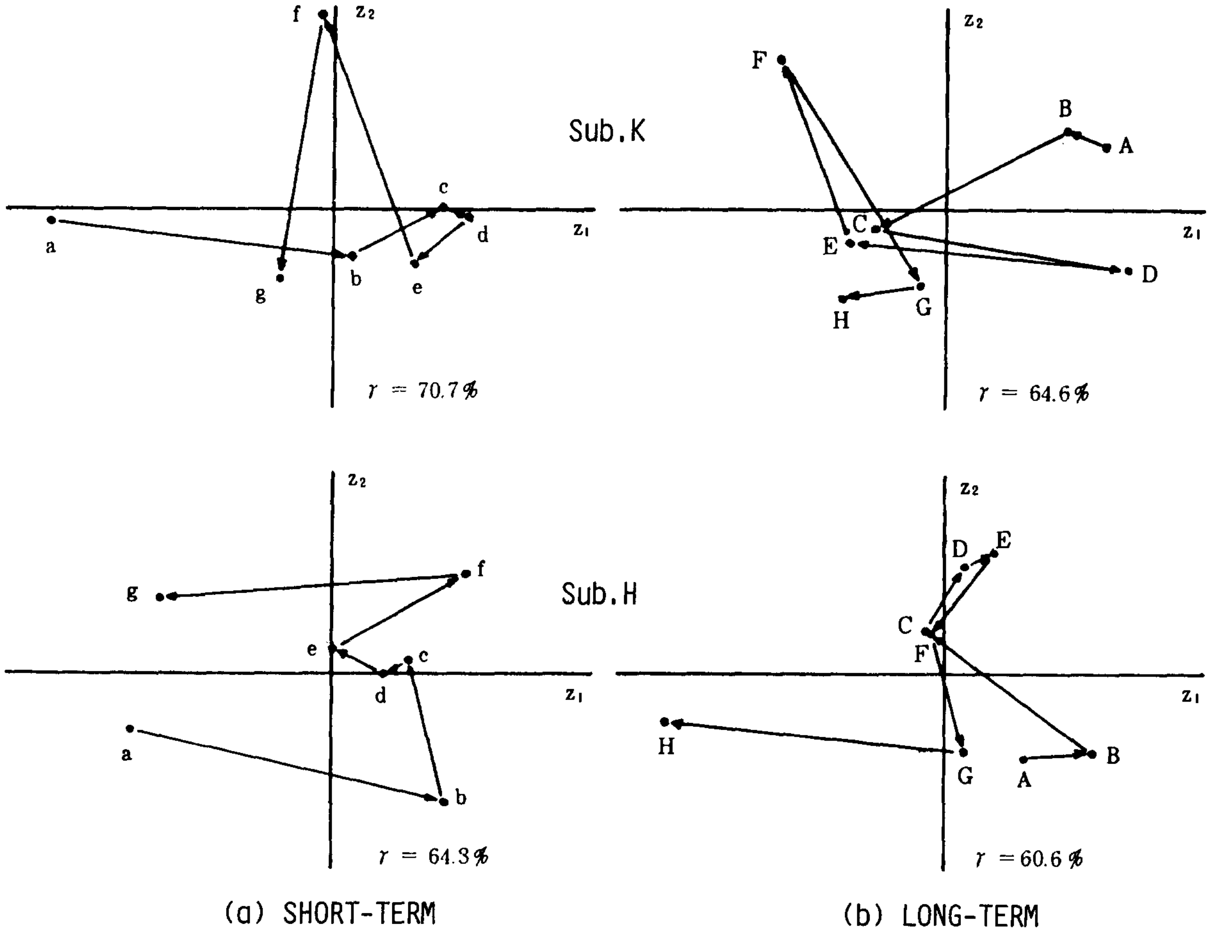
(b) LONG-TERM

Fig. 7. Time alignment of short-term and long-term speech data sets.

shows that the variance for the long-term data is nearly twice that for the short-term data, and that the increasing rate is much greater for the first-order coefficient.

In order to observe the long- and short-term variability of the LAS directly, a principal component analysis was applied to the long- and short-term cepstrum data sets for each speaker. The long-term data set consisted of samples recorded every three months over nearly two years, whereas the short-term data set consisted of samples recorded over a period of some five months with different intervals vaying between several days and three months, as shown in Fig. 7.

Examples of the utterances of two speakers projected onto the planes constructed by the first and the 2nd principal components are shown in Fig. 8 [11]. Similar results are indicated for utterances by seven other speakers. Results for the short-term data show that the utterances *b*, *c*, *d*



(a) SHORT-TERM

(b) LONG-TERM

Fig. 8. Short-term and long-term variation of a long-time averaged spectrum projected onto the planes constructed by the first and the second principal components. $z_1$: first component, $z_2$: second component, $\gamma$: the accumulated contribution rate for the first two principal components.

and $e$, which were recorded over ten days, are relatively close together. This means that the spectrum pattern is stable over short periods, whereas it presents large fluctuations over long periods. Results for the long-term data show that the spectrum fluctuates randomly for intervals of three months. It can also be seen that there is no uni-directional movement in either long-term or short-term data sets.

The variation amount was calculated for the cepstrum vector as a function of the time interval. Figure 9 shows the results averaged over nine male speakers, and these results indicate that the amount of variation increases monotonously as a function of the interval when the interval range is less than three months. On the other hand, it is nearly constant irrespective of the interval when it exceeds three months. This means that the variation of the LAS during the less-than-three-month period can be regarded as a multidimensional random walk, in which the distance from the original pattern statistically increases as a function of the time interval. However, when the pattern is sampled for intervals longer than three months, the variation can be characterized by a random variation around an expected pattern.

Under the simplified consideration of one-dimensional space, an expected value of the squared Euclidian distance, $d(\tau)$, can be related to the autocorrelation function, $\phi(\tau)$, as

$$d(\tau) = 2(\phi(0) - \phi(\tau)),$$

where $\tau$ indicates time delay. The dashed line in Fig. 9 indicates that the time function of the distance can be fitted by an exponential curve. This curve corresponds to the distance function derived from the correlation function for the signal consisting of the summation of two signal elements: a random signal passing through an RC filter with a time constant of 1.5 months and a small amplitude random signal which does not pass through the filter. The former element corresponds to the long-term variation of the pattern in which the variation depends on the interval length. The latter element corresponds to the fluctuation which is independent of the interval length.

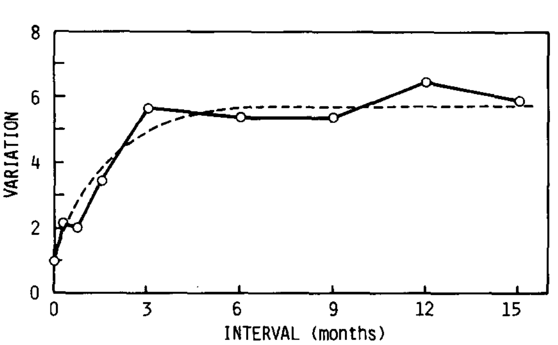


Fig. 9. Magnitude of variation for the cepstrum vector as a function of time interval averaged over nine male speakers. Dotted line shows the exponential curve fitting.

Since the LAS exhibits the above-mentioned variability, the accuracy of speaker recognition using the LAS largely decreases when the training utterances for each speaker are recorded over a short period such as 10 days, and the interval between the training and the input speech is three months or more, even if the cepstrum distance is weighted by a covariance matrix taking into consideration the variability of the training samples. However, when the training utterances are recorded over a long period such as 10 months, the error rate decreases to ½ and a 95% accuracy can be obtained, even if the interval between the training and input speech is three months.

Figure 10 shows the LAS variation for four sessions over eight months for two speakers. This figure also shows the corresponding spectral envelopes derived from the cepstrum coefficients for the LAS, which were weighted by the square roots of the inverse variances. Individual information which is stable over a long interval can be extracted from the LAS by the cepstrum expansion and the appropriate weighting. The spectral tilt is removed by de-weighting the first cepstrum coefficient with a large variance. This corresponds to the process for removing the overall spectral shape of glottal waves. This process is also effective for normalizing the variability of global transmission characteristics.

Figure 11 indicates the speaker effect (inter-to-intra-speaker variability ratio, $F$-ratio) for each cepstrum coefficient obtained through the one-factor analysis of variance using long-term speech data. As is shown, especially large speaker effects are demonstrated by the third and fourth elements corresponding to the spectral pattern which

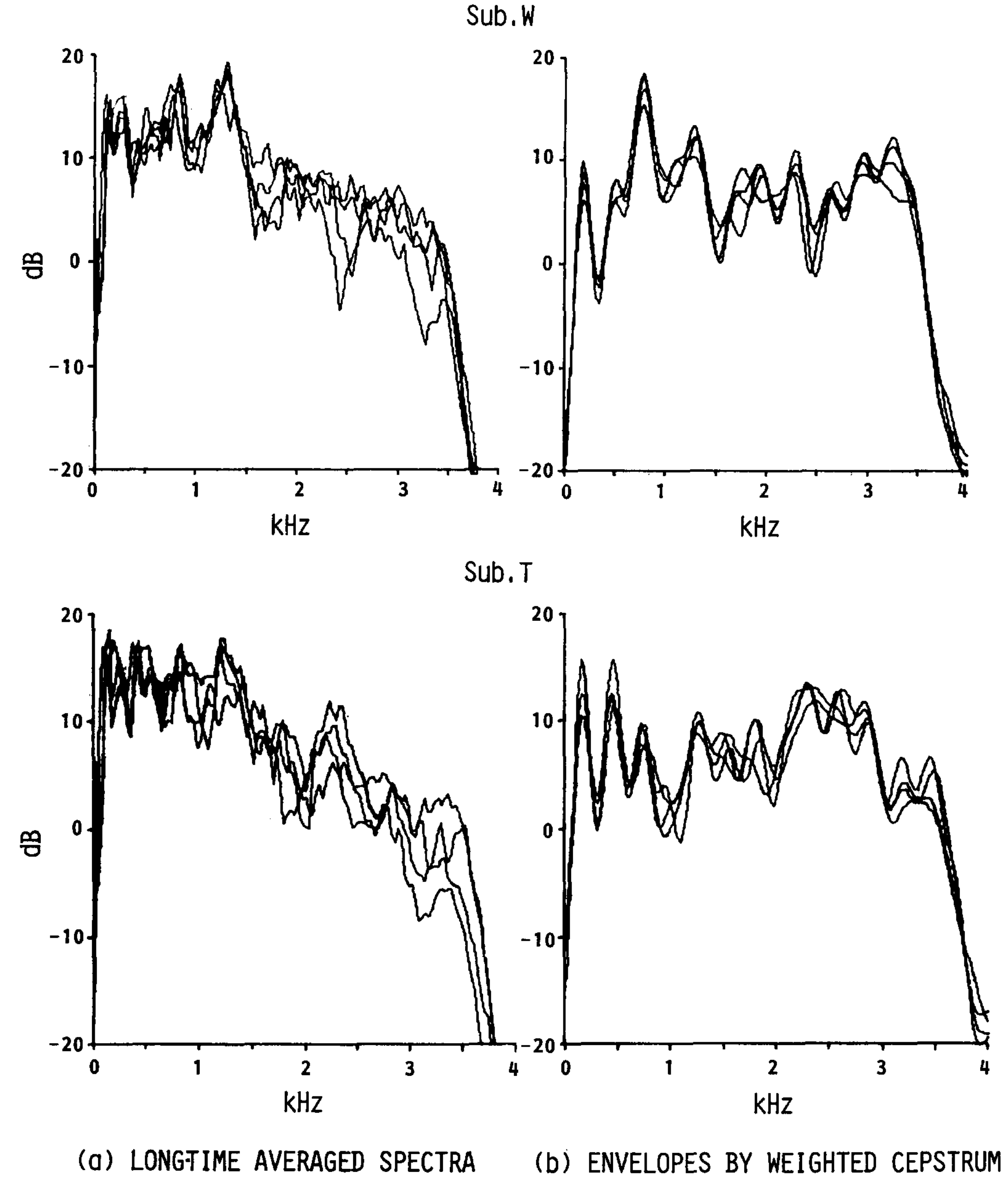


Fig. 10. Variation of the long-time averaged spectrum at four sessions over eight months, and corresponding spectral envelopes derived from cepstrum coefficients weighted by the square root of inverse variances.

repeats with a 2 to 2.7 kHz period along the frequency axis. This result corresponds to that described in the previous section, where the mean spectral level between 2.5 and 3.5 kHz is closely related to perceptual individuality.

### *4.2. Speaker recognition method based on explicit phoneme recognition*

The author [2] applied the method indicated in Fig. 12 as a trial example of the method shown in Fig. 2(b). Prior to recognition, multiple sets of

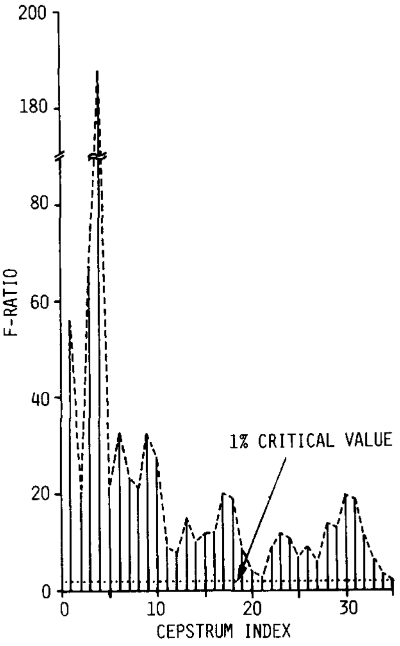


Fig. 11. Speaker effect (inter-to-intra-speaker variability ratio, $F$-ratio) obtained through the analysis of variance.

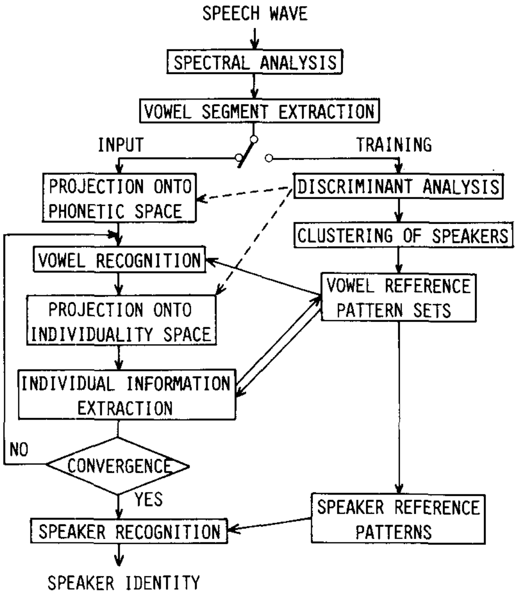


Fig. 12. Block diagram of the speaker recognition method based on the iteration of vowel recognition and individual information extraction.

vowel reference patterns for representing voice quality variation are constructed using a clustering technique. In the recognition stage, the vowel part of the input speech is recognized using these reference sets, and vowel dependent individual information is extracted on the basis of the vowel recognition results. The individual information extracted from all vowel segments is then combined to select a reference pattern set which best fits all the vowel segments in the input speech. Finally, recognition is attempted again, using the new reference set.

These iterations are effective in decreasing the early stage vowel recognition error influence produced by the variation of spectral parameters, and in stabilizing the recognition results. When the results converge, speaker recognition is performed using vowel-dependent reference patterns for each speaker based on the last vowel recognition results attained. Vowel recognition and individual information extraction are accomplished in different spaces produced by a discriminant analysis technique for emphasizing each characteristic. Experimental results show that the speaker identification and verification error rates are reduced to 1/3 of those obtained under the no-vowel recognition condition.

Matsumoto et al. [12] investigated a speaker recognition method based on the spectral space division into overlapping sub-spaces, each of which includes two or three phonemes ($i$'s) with similar interaction vectors $\gamma_{ij}$. With this method, short segments of input speech are automatically assigned to one of the sub-spaces, and projected into an individuality-emphasizing space constructed on the basis of the discriminant analysis applied to each sub-space. Speaker recognition performed in the projected spaces produces greater accuracy than attained with the method in which each phoneme is processed in an independent sub-space.

### 4.3. *Speaker recognition method based on implicit phoneme recognition*

Text-independent speaker recognition without explicit phoneme recognition can be realized by directly storing the vowel reference pattern set for each speaker. The reference pattern set which best fits the input speech is selected on the basis of a comparison of the input speech with all of these sets. Although this removal of the speakers' clustering represents a kind of simplification of the method shown in Fig. 12, it increases the number of calculations. Recently, various laboratories have attempted to use a modification of this method in which vowel recognition is not necessary. In this modification, codebooks constructed by vector quantization of continuous speech are used instead of vowel reference sets.

## 5. Text-dependent speaker recognition

### *5.1. Speaker recognition using statistical features of spectral parameters in word utterances*

The author et al. [13, 14] have used a method employing statistical features of spectral parameters extracted from spoken words as one of the text-dependent methods, i.e., the one in which words or sentences are predetermined. A block diagram of this method is shown in Fig. 13. Here, each speaker separately utters four kinds of words. Pitch frequency and first- to 10th-order cepstrum coefficients are extracted every 10 ms through the LPC analysis technique following spectral equalization. Thie spectral equalization is accomplished by first-order inverse filtering based on the autocorrelation coefficient averaged over each word length. A set of statistical features consisting of mean value, standard deviation, cross-correlation coefficients and Fourier cosine expansion coefficients are measured for the time functions of pitch frequency and cepstrum coefficients in the voiced portion of each word. A fixed reduced set, which is most effective in separating the speakers, is then selected and used for speaker recognition. This selection is based on the inter-to-intra-speaker variability ratio for each element, calculated through an analysis of variance over a series of training utterances.

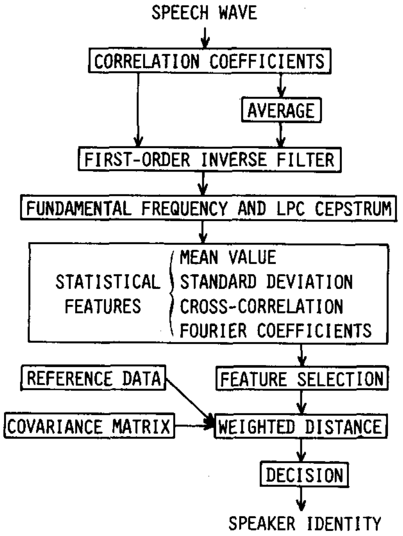


Fig. 13. Block diagram indicating the principal operation of the speaker recognition method using statistical features of spectral parameters extracted from a word utterance.

A recognition experiment was performed using speech utterances by nine male speakers recorded at 12 sessions over three years. Various sets of 12 samples over four sessions at three-monthly intervals were used as training samples for each speaker. The interval between training and input utterances was also set at three months. The number of impostors in speaker verification was 111. Experimental results demonstrated that a high recognition rate can be obtained by this method for both microphone and telephone speech.

An online speaker verification experiment using dialed-up telephone speech by 34 male and female speakers was conducted for six months. Although recognition accuracy was not high in the initial stages of the experiment due to the unreliability of the reference patterns, it increased with the increase in the number of experiments, reaching a final accuracy level of 97–98% at the stabilized stage.

One of the most crucial problems in speaker recognition system construction is the long-term intra-speaker variability of feature parameters and its effect on system performance. As mentioned in Subsection 4.1, the speech spectrum varies following a time interval, even if the same word is uttered by the same speaker, which thereby reduces the recognition accuracy. This problem, which is not an important matter in speech recognition, is crucial in speaker recognition since the training samples and input speech are always uttered at different sessions in the latter recognition and individual information is more detailed than the phonetic information.

The author [13,15] investigated methods for reducing the effects of long-term feature parameter variability on speaker recognition accruracy. The effectiveness of these methods for improving recognition accruracy was evaluated using the recognition system based on the statistical features of the spectrum derived from spoken words. In these investigations, the following methods were confirmed as effective for this purpose:

(1) The application of spectral equalization, i.e., the passing of the speech wave through a first- or second-order critical damping inverse filter which represents the overall pattern of the time-averaged spectrum for word or short-sentence speech. This process corresponds to the approximate removal of the glottal wave spectrum in speech or, in other words, to the preservation of vocal tract information only.

(2) The selection of stable feature parameters based both on the statistical evaluation of long-term speech data and on the combination of feature parameters extracted from different words or sentences.

(3) The construction of reference patterns and distance measures based on a set of long-term training samples.

(4) The renewal of the reference patterns at the appropriate time interval.

Experimental results confirm that when training samples are collected over a short period of 10 days and inverse filtering is not applied, the error rate largely increases when the interval between the training utterance and the input speech exceeds six weeks. Inverse filtering reduces the error rate to 1/3 however.

### 5.2. *Speaker recognition using spectral time series and dynamic features*

Figure 14 gives a block diagram of the speaker verification method which directly uses a spectral time series [16]. With this method, not only is the time series brought into time registration by the stored reference functions, but also a set of dynamic features is explicitly extracted and used for the recognition.

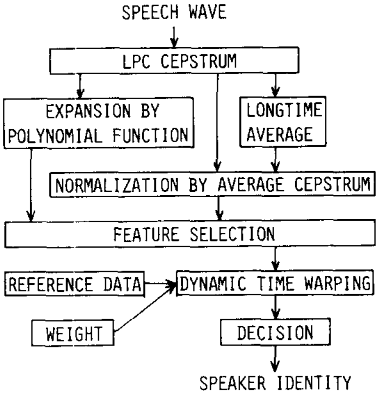


Fig. 14. Block diagram indicating the principal operation of the speaker recognition method using time series of cepstrum coefficients and their orthogonal polynomial coefficients.

Initially, 10 LPC cepstrum coefficients are extracted every 10 ms from short-sentence speech. These cepstrum coefficients are then averaged over the duration of the entire utterance, and the averaged values are subtracted from the cepstrum coefficients of every frame to compensate for the frequency-response distortions introduced by the transmission system, and to reduce long-term intra-speaker spectral variability. This procedure is similar to the spectral equalization described in the previous subsection. Time functions for the cepstrum coefficients are then expanded by an orthogonal polynomial representation over 90 ms intervals, which are shifted every 10 ms. The first- and second-order polynomial coefficients are thus obtained as the representations of dynamic characteristics. From the normalized cepstrum and polynomial coefficients, a set of 18 elements is selected which is most effective in separating speakers' overall distance-distribution. The time function of the set is brought into time registration with the reference template in order to calculate the distance between them. The overall distance is then compared with a threshold distance value for the verification decision. The threshold and reference template are updated every two weeks by using the distribution of inter-speaker distances.

Experimental results indicate that high verification accuracy can be obtained even if the reference and input utterances, both being telephone speech, are subjected to different transmission systems, such as ADPCM and LPC vocoder. An on-line experiment performed over a period of six months, using dialed-up telephone speech uttered by 60 male and 60 female speakers, also indicates the effectiveness of this system. This method of using dynamic features is effective not only for speaker recognition, but also for spoken word recognition [17].

Another speaker recognition method using dynamic spectral features has been investigated [18]. In this method, the slope of the dynamic time-warping function at the spectral transition interval in the input speech is examined during the course of time-warping matching between the input speech and the reference templates. This method is based on the principle that, with respect to the spectral transition parts, the slope of the warping function is usually maintained at around 45 degrees during intra-speaker comparisons. This is a result of the fact that the expansion and contraction of the speech duration usually happens at the spectrally stable intervals.

## 6. Discussion

The various research activities in Japan with respect to individual information embedded in speech waves and their application to speaker recognition have been reported upon in this paper. As one of the special focuses of these activities, the long-term intra-speaker variability of speech spectra has been investigated based on the databases collected over a very long period. In addition, the results are also reported [13] of a speaker-recognition experiment with an interval between training and input speech of as much as five years. Since the research on individual information is also very important for the realization of speaker-independent speech recognition, research in this area will become much more active in the near future.

Finally, three methods have been applied to automatic training in speech recognition for the purpose of constructing reference templates specifically adapted to each speaker:

(1) the approximate normalization of the individual difference of vocal tract length and glottal wave spectrum in the auto-correlation domain [19];

(2) efficient training [20] based on the correlation of individual features between different phonemes (cf. Section 2); and

(3) the improvement of the method described in Subsection 4.2, which uses multiple sets of reference patterns constructed by clustering techniques [21].

## References

[1] T. Sakai and K. Tabata, "Multivariate statistical analysis of VCV syllables", *Trans. IECE*, Vol. 56-D, 1973, pp. 63–70.

[2] S. Furui, "Research on individual information in speech waves", Ph. D. Thesis, Tokyo University, 1978.

[3] S. Furui, "Effects of transmission conditions on individual parameters in speech waves", *IECEJ National Meeting*, 1977.

[4] H. Matsumoto, et al., "Multidimensional representation of personal quality of vowels and its acoustical correlates", *IEEE Trans., Audio, Electro-acoustics*, Vol. 21, 1973, pp. 428–436.

[5] H. Kuwabara and K. Oogushi, "Psychological similarity of natural speech and its physical correlates", *Fall ASJ Meeting*, 1981.

[6] H. Kuwabara and K. Oogushi, "Acoustic characteristics of professional male announcers' speech", *Trans. IECE*, Vol. J66-A, 1983, pp. 545–552

[7] J.E. Sundberg, "Articulatory interpretation of the 'singing formant'", *J. Acoust. Soc. Amer.*,Vol. 55, 1974, pp. 838–844.

[8] K. Itoh and S. Saito, "Effects of acoustical feature parameters of speech on perceptual identification of speakers", *Trans. IECE*, Vol. J65-A, 1982, pp. 101–108.

[9] S. Furui and M. Akagi, "Perception of voice individuality and physical correlates", *Trans. Committee on Hearing Res., Acoust. Soc. Japan*, 1985.

[10] S. Furui, F. Itakura and S. Saito, "Talker recognition by longtime averaged speech spectrum", *Trans. IECE*, Vol. 55-A, 1972, pp. 549–556.

[11] S. Furui, F. Itakura and S. Saito, "Individual characteristics of the longtime averaged speech spectrum", *ECL Res. Development Report*, Vol. 23, 1974, pp. 1199–1210.

[12] H. Matsumoto, T. Sone and T. Nimura, "Text-independent speaker identification based on piecewise canonical discrimination analysis", *Trans. Comm. Electro-acoustics Res.*, 1975.

[13] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features", *IEEE Trans., Acoust., Speech, Signal Processing*, Vol. 29, 1981, pp. 342–350.

[14] S. Furui, "Speaker recognition by statistical features of

cepstral parameters", *Trans. IECE*, Vol. J65-A, 1982, pp. 183–190.

[15] S. Furui, "An analysis of long-term variation of feature parameters of speech and its application to talker recognition", *Trans. IECE*,Vol. 57-A, 1974, pp. 880–887.

[16] S. Furui, "Cepstral analysis technique for automatic speaker verification", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 29, 1981, pp. 254–272.

[17] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum", *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. 34, 1986, pp. 52–59.

[18] S. Saito and S. Furui, "Personal information in dynamic characteristics of speech spectra", *Proc. 4th IJCPR*, 1978, pp. 1014–1018.

[19] S. Furui, "Learning and normalization of the talker differences in the recognition of spoken words", *Trans. Comm. Speech Res., Acoust. Soc. Japan.*, 1975

[20] S. Furui, "A training procedure for isolated word recognition systems", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 28, 1980, pp. 129–136.

[21] M. Sugiyama and M. Kohda, "An unsupervised speaker adaptation technique of vowel templates using speech recognition results", *Trans. Comm. Speech Res., Acoust. Soc. Japan.*, 1985.