ELSEVIER

Pattern Recognition Letters 18 (1997) 859–872

Pattern Recognition Letters

Recent advances in speaker recognition

Sadaoki Furui [1]

Tokyo Institute of Technology, 2-12-1, O-okayama, Meguro-ku, Tokyo 152, Japan

Abstract

This paper introduces recent advances in speaker recognition technology. The first part discusses general topics and issues. The second part is devoted to a discussion of more specific topics of recent interest that have led to interesting new approaches and techniques. They include VQ- and ergodic-HMM-based text-independent recognition methods, a text-prompted recognition method, parameter/distance normalization and model adaptation techniques, and methods of updating models and a priori thresholds in speaker verification. Although many recent advances and successes have been achieved in speaker recognition, there are still many problems for which good solutions remain to be found. The last part of this paper describes 16 open questions about speaker recognition. The paper concludes with a short discussion assessing the current status and future possibilities. © 1997 Elsevier Science B.V.

Keywords: Speaker recognition; Speaker verification; Speker identification; Text-prompted method; HMM; Likelihood normalization

1. Principles of speaker recognition

1.1. What is speaker recognition?

Speaker recognition is the process of automatically recognizing who is speaking by using speaker-specific information included in speech waves (Doddington, 1985; Furui, 1986, 1989, 1991a,b, 1994; O'Shaugnessy, 1986; Rosenberg and Soong, 1991). This technique can be used to verify the identity claimed by people accessing systems; that is, it enables access control of various services by voice. Applicable services include voice dialing, banking over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information, and remote access of computers. Another important application of speaker recognition technology is its use for forensic purposes (Kunzel, 1994).

1.2. Classification of speaker recognition

Speaker recognition can be classified into speaker identification and speaker verification. Speaker identification is the process of determining from which of the registered speakers a given utterance comes. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. Most of the applications in which voice is used to confirm the identity claim of a speaker are classified as speaker verification. The fundamental difference between identification and verification is the number of decision alternatives. In identification, the number of decision alternatives is equal to the size of the population, whereas in verification there are only

[1] Email: furui@cs.titech.ac.jp.

PII S0167-8655(97)00073-1

two choices, accept or reject, regardless of the population size. Therefore, speaker identification performance decreases as the size of the population increases, whereas speaker verification performance approaches a constant, independent of the size of the population, unless the distribution of physical characteristics of speakers is extremely biased.

There is also the case called "open set" identification, in which a reference model for the unknown speaker may not exist. In this case, an additional decision alternative, "the unknown does not match any of the models", is required. In either verification or identification, an additional threshold test can be applied to determine whether the match is close enough to accept the decision or ask for a new trial.

Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former require the speaker to provide utterances of key words or sentences that are the same text for both training and recognition, whereas the latter do not rely on a specific text being spoken. The text-dependent methods are usually based on template-matching techniques in which the time axes of an input speech sample and each reference template or reference model of the registered speakers are aligned, and the similarity between them is accumulated from the beginning to the end of the utterance (Furui, 1981; Naik et al., 1989; Rosenberg et al., 1991; Zheng and Yuan, 1988). Since this method can directly exploit voice individuality associated with each phoneme or syllable, it generally achieves higher recognition performance than the text-independent method.

However, there are several applications, such as forensic and surveillance applications, in which predetermined key words cannot be used. Moreover, human beings can recognize speakers irrespective of the content of the utterance. Therefore, text-independent methods have recently attracted more attention. Another advantage of text-independent recognition is that it can be done sequentially, until a desired significance level is reached, without the annoyance of the speaker having to repeat the key words again and again.

Both text-dependent and independent methods have a serious weakness. That is, these systems can easily be defeated, because someone who plays back the recorded voice of a registered speaker uttering key words or sentences into the microphone can be accepted as the registered speaker. To cope with this problem, some methods use a small set of words, such as digits, as key words, and each user is prompted to utter a given sequence of key words that is randomly chosen every time the system is used (Higgins et al., 1991; Rosenberg et al., 1991). Yet even this method is not reliable enough, since it can be defeated with advanced electronic recording equipment that can reproduce key words in a requested order. Therefore, a text-prompted speaker recognition method has recently been proposed. (See Section 2.3.)

1.3. Basic structures of speaker recognition systems

In the speaker identification task, a speech utterance from an unknown speaker is analyzed and compared with speech models of known speakers. The unknown speaker is identified as the speaker whose model best matches the input utterance. In speaker verification, an identity claim is made by an unknown speaker, and an utterance of this unknown speaker is compared with the model for the speaker whose identity is claimed. If the match is good enough, that is, above a threshold, the identity claim is accepted. A high threshold makes it difficult for impostors to be accepted by the system, but at the risk of falsely rejecting valid users. Conversely, a low threshold enables valid users to be accepted consistently, but at the risk of accepting impostors. To set the threshold at the desired level of customer rejection and impostor acceptance, it is necessary to know the distribution of customer and impostor scores.

The effectiveness of speaker-verification systems can be evaluated by using the receiver operating characteristics (ROC) curve adopted from psychophysics. The ROC curve is obtained by assigning two probabilities, the probability of correct acceptance and the probability of incorrect acceptance, to the vertical and horizontal axes respectively, and varying the decision threshold (Furui, 1989). The equal-error rate (EER) is a commonly accepted overall measure of system performance. It corresponds to the threshold at which the false acceptance rate is equal to the false rejection rate.

1.4. Feature parameters

Speaker identity is correlated with the physiological and behavioral characteristics of the speech production system for each speaker. These characteristics exist both in the spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics) of speech. Although it is impossible to separate these kinds of characteristics, and many voice characteristics are difficult to measure explicitly, many characteristics are captured implicitly by various signal measurements. Signal measurements such as short-term and long-term spectra and overall energy are easy to obtain. These measurements provide the means for effectively discriminating among speakers. Fundamental frequency can also be used to recognize speakers if it can be extracted reliably (Atal, 1972; Carey et al., 1996; Matsui and Furui, 1990).

Currently, the most commonly used short-term spectral measurements are LPC-derived cepstral coefficients and their regression coefficients (Furui, 1981). A spectral envelope reconstructed from a truncated set of cepstral coefficients is much smoother than one reconstructed from LPC coefficients, and hence provides a stabler representation from one repetition to another of a particular speaker's utterances. As for the regression coefficients, typically, the first- and second-order coefficients, that is, derivatives of the time functions of cepstral coefficients, are extracted at every frame period to represent spectral dynamics. They are respectively called the delta-cepstral and delta-delta-cepstral coefficients.

1.5. Text-dependent speaker recognition methods

Text-dependent speaker recognition methods can be classified into DTW (dynamic time warping) or HMM (hidden Markov model) based methods.

1.5.1. DTW-based methods

A typical approach to text-dependent speaker recognition is the spectral template matching approach (Furui, 1981). In this approach, each utterance is represented by a sequence of feature vectors, generally, short-term spectral feature vectors, and the trial-to-trial timing variation of utterances of the same text is normalized by aligning the analyzed feature vector sequence of a test utterance to the template feature vector sequence using a DTW algorithm. The overall distance between the test utterance and the template is used for recognition decision.

1.5.2. HMM-based methods

An HMM can efficiently model the statistical variation in spectral features. Therefore, HMM-based methods have achieved significantly better recognition accuracies than DTW-based methods (Naik et al., 1989; Rosenberg et al., 1991; Zheng and Yuan, 1988).

A speaker verification system based on characterizing the utterances as sequences of subword units represented by HMMs has been introduced and tested (Rosenberg et al., 1990a,b). Two types of subword units, phone-like units (PLUs) and acoustic segment units (ASUs), have been studied. PLUs are based on phonetic transcriptions of spoken utterances and ASUs are extracted directly from the acoustic signal without using any linguistic knowledge. The results of experiments using isolated digit utterances show only small differences in performance between PLU- and ASU-based representations.

1.6. Text-independent speaker recognition methods

In text-independent speaker recognition, the words or sentences used in recognition trials generally cannot be predicted. Since it is impossible to model or match speech events at the word or sentence level, the following three kinds of methods shown in Fig. 1 have been actively investigated (Furui, 1986).

1.6.1. Long-term-statistics-based methods

As text-independent features, long-term sample statistics of various spectral features, such as the mean and variance of spectral features over a series of utterances, have been used (Furui et al., 1972; Markel et al., 1977; Markel and Davi, 1979) (Fig. 1(a)). However, long-term spectral averages are extreme condensations of the spectral characteristics of a speaker's utterances and, as such, lack the discriminating power included in the sequences of short-term spectral features used as models in text-dependent methods. In one of the trials using the long-term

averaged spectrum (Furui et al., 1972), the effect of session-to-session variability was reduced by introducing a weighted cepstral distance measure.

Studies on using statistical dynamic features have also been reported. Montacie et al. (1992) applied a multivariate auto-regression (MAR) model to the time series of cepstral vectors to characterize speakers, and reported good speaker recognition results. Griffin et al. (1994) studied distance measures for the MAR-based method, and reported that when ten sentences were used for training and one sentence was used for testing, identification and verification rates were almost the same as those obtained by an HMM-based method.

It was also reported that the optimum order of the MAR model was 2 or 3, and that distance normalization using a posteriori probability was essential to obtain good results in speaker verification.

1.6.2. VQ-based methods

A set of short-term training feature vectors of a speaker can be used directly to represent the essential characteristics of that speaker. However, such a direct representation is impractical when the number of training vectors is large, since the memory and amount of computation required become prohibitively large. Therefore, attempts have been made to find efficient ways of compressing the training data using vector quantization (VQ) techniques.

In this method, VQ codebooks, consisting of a small number of representative feature vectors, are used as an efficient means of characterizing speaker-specific features (Li and Wrench Jr, 1983; Matsui and Furui, 1990; Matsui and Furui, 1991; Rosenberg and Soong, 1987; Shikano, 1985; Soong et al., 1987). A speaker-specific codebook is generated by clustering the training feature vectors of each speaker. In the recognition stage, an input utterance is vector-quantized by using the codebook of each reference speaker; the VQ distortion accumulated over the entire input utterance is used for making the recognition determination (Fig. 1(b)).

1.6.3. Ergodic-HMM-based methods

The basic structure is the same as the VQ-based method (Fig. 1(b)), but in this method an ergodic

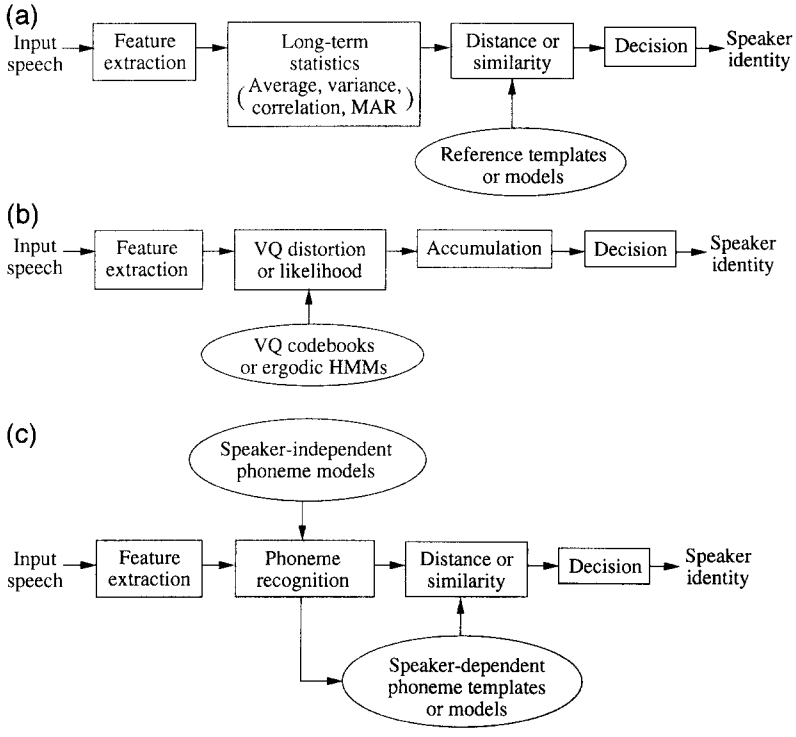

Fig. 1. Basic structures of text-independent speaker recognition methods.

HMM is used instead of a VQ codebook. Over a long timescale, the temporal variation in speech signal parameters is represented by stochastic Markovian transitions between states. Poritz (1982) proposed using a five-state ergodic HMM (i.e., all possible transitions between states are allowed) to classify speech segments into one of the broad phonetic categories corresponding to the HMM states. A linear predictive HMM was adopted to characterize the output probability function. He characterized the automatically obtained categories as strong voicing, silence, nasal/liquid, stop burst/post silence, and frication.

Tishby (1991) extended Poritz's work to the richer class of mixture autoregressive (AR) HMMs. In these models, the states are described as a linear combination (mixture) of AR sources. It can be shown that mixture models are equivalent to a larger HMM with simple states, together with additional constraints on the possible transitions between states.

1.6.4. Speech-recognition-based methods

The VQ- and HMM-based methods can be regarded as methods that use phoneme-class-dependent speaker characteristics in short-term spectral features through implicit phoneme-class recognition. In other words, phoneme-classes and speakers are simultaneously recognized in these methods. On the other hand, in the speech-recognition-based methods (Fig. 1(c)), phonemes or phoneme-classes are explicitly recognized, and then each phoneme (-class) segment in the input speech is compared with speaker models or templates corresponding to that phoneme (-class).

Savic and Gupta (1990) used a five-state ergodic linear predictive HMM for broad phonetic categorization. In their method, after frames that belong to particular phonetic categories have been identified, feature selection is performed. In the training phase, reference templates are generated and verification thresholds are computed for each phonetic category. In the verification phase, after phonetic categorization, a comparison with the reference template for each particular category provides a verification score for that category. The final verification score is a weighted linear combination of the scores for each category. The weights are chosen to reflect the effectiveness of particular categories of phonemes in discriminating between speakers and are adjusted to maximize the verification performance. Experimental results showed that verification accuracy can be considerably improved by this category-dependent weighted linear combination method. Broad phonetic categorization can also be implemented by a speaker-specific hierarchical classifier instead of by an HMM, and the effectiveness of this approach has also been confirmed (Eatock and Mason, 1990).

Recently, speaker verification through large vocabulary continuous speech recognition has been investigated. Details are given in Section 2.2.

2. Recent advances

2.1. VQ- and HMM-based text-independent methods

Matsui and Furui (1990, 1991) tried a method using a VQ-codebook for long feature vectors consisting of instantaneous and transitional features calculated for both cepstral coefficients and fundamental frequency. Since the fundamental frequency cannot be extracted from unvoiced speech, there are two separate codebooks for voiced and unvoiced speech for each speaker. A new distance measure was introduced to take into account the intra- and inter-speaker variability and to deal with the outlier problem in the distribution of feature vectors. The outlier vectors correspond to intersession spectral variation and to the difference between phonetic content of the training texts and the test utterances. Experimental results confirmed high recognition accuracies even when the codebooks for each speaker were made using training utterances recorded in a single session and the time difference between training and testing was more than three months. It was also confirmed that, although the fundamental frequency achieved only a low recognition rate by itself, the recognition accuracy was greatly improved by combining the fundamental frequency with spectral envelope features.

In contrast with the memoryless VQ-based method, non-memoryless source coding algorithms have also been studied using a segment (matrix) quantization technique (Juang and Soong, 1990; Sugiyama, 1988). The advantage of a segment quantization codebook over a VQ codebook representation is its characterization of the sequential nature of speech events. Higgins and Wohlford (1986) pro-

posed a segment modeling procedure for constructing a set of representative time normalized segments, which they called "filler templates". The procedure, a combination of K-means clustering and dynamic programming time alignment, provided a means for handling temporal variation.

Matsui and Furui (1992) compared the VQ-based method with the discrete/continuous ergodic HMM-based method, particularly from the viewpoint of robustness against utterance variations. They found that the continuous ergodic HMM method is far superior to the discrete ergodic HMM method and that the continuous ergodic HMM method is as robust as the VQ-based method when enough training data is available. However, when little data is available, the VQ-based method is more robust than the continuous HMM method. They investigated speaker identification rates using the continuous HMM as a function of the number of states and mixtures. It was shown that the speaker recognition rates are strongly correlated with the total number of mixtures, irrespective of the number of states. This means that the information on transitions between different states is ineffective for text-independent speaker recognition.

The case of a single-state continuous ergodic HMM corresponds to the technique based on the maximum likelihood estimation of a Gaussian-mixture model representation investigated by Rose and Reynolds (1990). Furthermore, the VQ-based method can be regarded as a special (degenerate) case of a single-state HMM with a distortion measure being used as the observation probability. Gaussian mixtures are noted for their robustness as a parametric model and their ability to form smooth estimates of rather arbitrary underlying densities.

The ASU-based speaker verification method described in Section 1.5.2 has also been tested in the text-independent mode (Rosenberg et al., 1990a,b). It has been shown that this approach can be extended to large vocabularies and continuous speech.

2.2. *Speech-recognition-based speaker recognition*

Gauvain et al. (1995) investigated a statistical modeling approach, where each speaker was viewed as a source of phonemes, modeled by a fully connected Markov chain. Maximum a posteriori (MAP) estimation was used to generate speaker-specific models from a set of speaker-independent seed models. The lexical and syntactic structures of the language were approximated by local phonotactic constraints. The unknown speech is recognized by all of the speakers' models in parallel, and the hypothesized identity is that associated with the model set having the highest likelihood.

Since phonemes and speakers are simultaneously recognized by using speaker-specific Markov chains, this method can be considered as an extension of the ergodic-HMM-based method. The experimental results using the BREF corpus showed that this method clearly out-performed a simpler Gaussian mixture (single-state HMM) model. It was also found that text-independent and text-dependent verification EERs were about the same.

Rosenberg et al. have been testing a speaker verification system using 4-digit phrases under field conditions of a banking application (Rosenberg et al., 1991; Setlur and Jacobs, 1995). In this system, input speech is segmented into individual digits using a speaker-independent HMM. The frames within the word boundaries for a digit are compared with the corresponding speaker-specific HMM digit model and the Viterbi likelihood score is computed. This is done for each of the digits making up the input utterance. The verification score is defined to be the average normalized log-likelihood score over all the digits in the utterance.

Newman et al. (1996) used a large vocabulary speech recognition system for speaker verification. A set of speaker-independent phoneme models were adapted to each speaker. The speaker verification consisted of two stages. First, speaker-independent speech recognition was run on each of the test utterances to obtain phoneme segmentation. In the second stage, the segments were scored against the adapted models for a particular target speaker. The scores were normalized by those with speaker-independent models. The system was evaluated using the 1995 NIST-administered speaker verification database, which consists of data taken from the Switchboard corpus. The results showed that this method could not out-perform Gaussian mixture models.

2.3. Text-prompted speaker recognition

How can we prevent speaker verification systems from being defeated by a recorded voice? Another problem is that people often do not like text-dependent systems because they do not like to utter their identification number, such as their social security number, within the hearing of other people. To cope with these problems, a text-prompted speaker recognition method has been proposed.

In this method, key sentences are completely changed every time (Matsui and Furui, 1993, 1994a,b). The system accepts the input utterance only when it determines that the registered speaker uttered the prompted sentence. Because the vocabulary is unlimited, prospective impostors cannot know in advance the sentence they will be prompted to say. This method can not only accurately recognize speakers, but can also reject an utterance whose text differs from the prompted text, even if it is uttered by a registered speaker. Thus, a recorded and played back voice can be correctly rejected.

This method uses speaker-specific phoneme models as basic acoustic units. One of the major issues in this method is how to properly create these speaker-specific phoneme models when using training utterances of a limited size. The phoneme models are represented by Gaussian-mixture continuous HMMs or tied-mixture HMMs, and they are made by adapting speaker-independent phoneme models to each speaker's voice. Since the text of training utterances is known, these utterances can be modeled as the concatenation of phoneme models, and these models can be automatically adapted by an iterative algorithm.

In the recognition stage, the system concatenates the phoneme models of each registered speaker to create a sentence HMM, according to the prompted text. Then the likelihood of input speech against the sentence model is calculated and used for the speaker recognition determination. If the likelihood of both speaker and text is high enough, the speaker is accepted as the claimed speaker. Experimental results gave a speaker and text verification rate of 99.4% when the adaptation method for tied-mixture-based phoneme models and the likelihood normalization method described in the Section 2.4.2 were used.

2.4. Normalization and adaptation techniques

How can we normalize the intra-speaker variation of likelihood (similarity) values in speaker verification? The most significant factor affecting automatic speaker recognition performance is variation in signal characteristics from trial to trial (intersession variability or variability over time). Variations arise from the speaker him/herself, from differences in recording and transmission conditions, and from noise. Speakers cannot repeat an utterance precisely the same way from trial to trial. It is well known that samples of the same utterance recorded in one session are much more highly correlated than tokens recorded in separate sessions. There are also long term trends in voices (Furui et al., 1972; Furui, 1974).

It is important for speaker recognition systems to accommodate these variations. Adaptation of the reference model as well as the verification threshold for each speaker is indispensable to maintain a high recognition accuracy for a long period. In order to compensate for the variations, two types of normalization techniques have been tried – one in the parameter domain, and the other in the distance/similarity domain. The latter technique uses the likelihood ratio or a posteriori probability. To adapt HMMs for noisy conditions, the HMM composition (PMC: parallel model combination) method has been successfully tried.

2.4.1. Parameter-domain normalization

As one typical normalization technique in the parameter domain, spectral equalization, the so-called "blind equalization" method, has been confirmed to be effective in reducing linear channel effects and long-term spectral variation (Atal, 1974; Furui, 1981). This method is especially effective for text-dependent speaker recognition applications using sufficiently long utterances. In this method, cepstral coefficients are averaged over the duration of an entire utterance, and the averaged values are subtracted from the cepstral coefficients of each frame. This method can compensate fairly well for additive variation in the log spectral domain. However, it unavoidably removes some text-dependent and speaker-specific features, so it is inappropriate for short utterances in speaker recognition applications.

It was shown that time derivatives of cepstral coefficients (delta-cepstral coefficients) are resistant to linear channel mismatch between training and testing (Furui, 1981; Soong and Rosenberg, 1988).

2.4.2. Likelihood normalization

Higgins et al. (1991) proposed a normalization method for distance (similarity or likelihood) values that uses a likelihood ratio:

$$\log L(X) = \log p(X \mid S = S_c) - \log p(X \mid S \neq S_c). \quad (1)$$

The likelihood ratio is the ratio of the conditional probability of the observed measurements of the utterance given the claimed identity is correct to the conditional probability of the observed measurements given the speaker is an impostor. Generally, a positive value of $\log L$ indicates a valid claim, whereas a negative value indicates an impostor. We call the second term on the right-hand side of Eq. (1) the normalization term.

The density at point X for all speakers other than true speaker S can be dominated by the density for the nearest reference speaker, if we assume that the set of reference speakers is representative of all speakers. This means that likelihood ratio normalization approximates optimal scoring in Bayes' sense. This normalization method is, however, unrealistic because, even if only the nearest reference speaker is used, conditional probabilities must be calculated for all the reference speakers, which costs a lot. Therefore, a set of speakers, "cohort speakers", has been chosen for calculating the normalization term of Eq. (1). Higgins et al. proposed using speakers that are representative of the population near the claimed speaker.

An experiment in which the size of the cohort speaker set was varied from 1 to 5 showed that speaker verification performance increases as a function of the cohort size, and that the use of normalization significantly compensates for the degradation obtained by comparing verification utterances recorded using an electret microphone with models constructed from training utterances recorded with a carbon button microphone (Rosenberg, 1992).

Matsui and Furui (1993, 1994a,b) proposed a normalization method based on a posteriori probability:

$$\log L(X) = \log p(X \mid S = S_c) - \log \sum_{S \in Ref} p(X \mid S). \quad (2)$$

The difference between the normalization method based on the likelihood ratio and that based on a posteriori probability is whether or not the claimed speaker is included in the impostor speaker set for normalization; the cohort speaker set in the likelihood-ratio-based method does not include the claimed speaker, whereas the normalization term for the a posteriori-probability-based method is calculated by using a set of speakers including the claimed speaker. Experimental results indicate that both normalization methods almost equally improve speaker separability and reduce the need for speaker-dependent or text-dependent thresholding, compared with scoring using only the model of the claimed speaker (Matsui and Furui, 1994a,b; Rosenberg, 1992).

The normalization method using the cohort speakers that are representative of the population near the claimed speaker is expected to increase the selectivity of the algorithm against voices similar to the claimed speaker. However, this method has a serious problem that it is vulnerable to attack by impostors of the opposite gender. Since the cohorts generally model only same-gender speakers, the probability of opposite-gender impostor speech is not well modeled and the likelihood ratio is based on the tails of distributions, which gives rise to unreliable values. Another way of choosing the cohort speaker set is to use speakers who are typical of the general population. Reynolds (1994) reported that a randomly selected, gender-balanced background speaker population outperformed a population near the claimed speaker.

Carey and Parris (1992) proposed a method in which the normalization term is approximated by the likelihood for a world model representing the population in general. This method has an advantage that the computational cost for calculating the normalization term is much smaller than the original method since it does not need to sum the likelihood values for cohort speakers. Matsui and Furui (1994a,b)

recently proposed a new method based on tied-mixture HMMs in which the world model is made as a pooled mixture model representing the parameter distribution for all the registered speakers. This model is created by averaging together the mixture-weighting factors of each reference speaker calculated using speaker-independent mixture distributions. Therefore the pooled model can be easily updated when a new speaker is added as a reference speaker. In addition, this method has been confirmed to give much better results than either of the original normalization methods.

Since these normalization methods neglect the absolute deviation between the claimed speaker's model and the input speech, they cannot differentiate highly dissimilar speakers. Higgins et al. (1991) reported that a multilayer network decision algorithm makes effective use of the relative and absolute scores obtained from the matching algorithm.

2.4.3. HMM adaptation for noisy conditions

How can we improve the robustness of speaker recognition techniques against noisy speech or speech distorted by a telephone? The robustness issue is crucial in real-world applications. Will speaker recognition technology ever be capable of performing with high reliability under adverse real-world conditions (noisy telephone lines, limited training data, etc.)? Only a few papers have addressed this problem in speaker recognition.

Rose et al. (1994) applied the HMM composition (PMC: parallel model combination) method (Gales and Young, 1993; Martin et al., 1993) to speaker identification under noisy conditions. The HMM composition is a technique to combine a clean speech HMM and a background noise HMM to create a noise-added speech HMM. They can be optimally combined by using the expectation-maximization (EM) algorithm. Experimental results show that this method is highly effective at recognizing speech with additive noise. In this method, the signal-to-noise ratio (SNR) of input speech is assumed to be similar to that of training speech or to be given. However, the SNR is variable in real situations and it is also difficult to measure especially when the noise is non-stationary. Matsui and Furui (1996a,b) proposed a method in which several noise-added HMMs with various SNRs were created and the HMM that had the highest likelihood value for the input speech was selected. A speaker decision was made using the likelihood value corresponding to the selected model. Experimental application of this method to text-independent speaker identification and verification in various kinds of noisy environments demonstrated considerable improvement in speaker recognition.

2.5. Updating models and a priori threshold for speaker verification

How do we deal with long-term variability in people's voices? How should we update the speaker models to cope with the gradual changes in people's voices? Since we cannot ask every user to utter many utterances at many different sessions in real situations, it is necessary to build each speaker model based on a small amount of data collected at a few sessions, and then the model must be updated using speech data collected when the system is used. How do we adequately retrain the models?

How should we set the a priori decision threshold for speaker verification? In most laboratory speaker recognition experiments, the threshold is set a posteriori so that the equal error rate (EER) is achieved. Since the threshold cannot be set a posteriori in real situations, we have to have reasonable ways to set the threshold before verification. It must be set according to the importance of the two errors, which depends on the application.

These two problems are highly related each other. Furui (1981) proposed methods for updating reference templates and the threshold in DTW-based speaker verification. An optimum threshold was estimated based on the distribution of overall distances between each speaker's reference template and a set of utterances of other speakers (interspeaker distances). The interspeaker distance distribution was approximated by a normal distribution, and the threshold was calculated by the linear combination of its mean value and standard deviation. The intraspeaker distance distribution was not taken into account in the calculation, mainly because it is difficult to obtain stable estimates of the intraspeaker distance distribution for small numbers of training utterances. It is easy to estimate interspeaker distance distributions by cross comparison of the training utterances between different speakers. The threshold

was updated at the same time as the reference templates. The reference template for each speaker was updated by averaging new utterances and the present template after time registration.

Matsui and Furui (1996a,b) extended these methods and applied them to text-independent and text-prompted speaker verification using HMMs. HMM parameters for each speaker were updated by using present parameters and a limited amount of data for adaptation so that they were close to the parameter values given by maximum likelihood estimation using all the past and present data. The thresholds for speaker verification were updated according to Eq. (3). The threshold converges from a threshold value that has a higher false acceptance (FA) rate to the EER threshold value for the samples for updating the model as the updating of the model proceeds.

$$\tilde{\phi} = \omega\phi_1 + (1-\omega)\phi_0, \tag{3}$$

$$\omega = \frac{2}{1+\exp(a\cdot k)}. \tag{4}$$

Here, ϕ_0 is the e.e.t. for the samples for updating the model; ϕ_1 is the threshold which determines the upper bound of the FA rate (Fig. 2). In the experiment, ϕ_1 was set to where the FA rate was slightly higher that the EER. The ω is a parameter for controlling the convergence of the threshold, and for instance, defined as (4) with k being the number of sessions for updating the model and a being an experimental parameter. Evaluation of the performance of the two methods using 20 male speakers showed that the verification error rate was about 40% of that without updating in text-independent experiments, and about 80% in text-prompted ones.

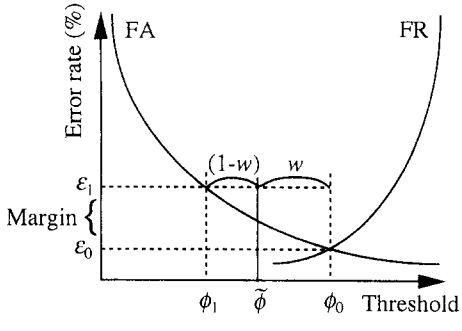

Fig. 2. Method of updating the speaker verification threshold.

Setlur and Jacobs (1995) showed that speaker verification performance was improved by constraining the individual feature variances for all models to a fixed set of values when the size of training data was small. This is because the model variance estimates derived from the K-means clustering approach using small data are poor and result in erratic performance. The fixed set of variances was derived by averaging the model variances of the speaker-independent model across all states and mixtures for each feature. The decision threshold was determined a priori by using an independent impostor set and a target false acceptance rate.

3. Open questions about speaker recognition

Although many recent advances and successes have been achieved as described in the previous sections, there are still many problems for which good solutions remain to be found. Sixteen major problems are discussed below.

(1) *How can human beings correctly recognize speakers?*

Research on perceptual voice individuality has been conducted for many years, but what we have learned about the hearing capability of human beings is very limited. Which acoustic parameters do listeners utilize in making their judgments? One of the results found by experiment is that segmental (spectral envelope) information plays a more important role than supra-segmental (pitch and energy) information. It is very difficult to give an answer to a question like "which performs better, a human being or a computer". The answer very much depends on the conditions, such as the number of speakers, their familiarity to the listeners, noise and distortion, and the time difference between training and testing.

(2) *Is it useful to study the mechanism of speaker recognition by human beings?*

Do auditory models help improve speaker recognition performance? An example of how knowledge of human hearing characteristics led to a new idea for automatic speaker recognition was the creation of delta-cepstral parameters (polynomial expansion coefficients of cepstral parameters). These parameters were proposed based on the experimental results that

our hearing systems are very sensitive to spectral transition. Although the number of such examples is very small, we expect advancements in our understanding of the mechanism of speech perception to help create new engineering ideas in the future.

(3) *Is it useful to study the physiological mechanism of speech production to get new ideas for speaker recognition?*

How is identity encoded and is this worth pursuing for automatic speaker recognition? How are perceptually identical voices produced by our vocal systems?

(4) *What feature parameters are appropriate for speaker recognition?*

The most common parameters we use are cepstral and delta-cepstral parameters. These are exactly the same as those used in speech recognition. Is it worth searching for a new set of feature parameters that is more appropriate to speaker recognition than the present set?

(5) *How can we model the relatively macro-transitional features covering the interval between 200 to 300 ms?*

The delta-cepstral parameters usually represent the transitional features covering the interval between 50 to 100 ms. The macro-transitional features covering the interval between 200 to 300 ms correspond to phoneme-to-phoneme and syllable-to-syllable transitions. We do not know how to represent these transitional features for speaker recognition. As described in Section 2.1, it was found that transitional information between different states in HMM is ineffective for text-independent speaker recognition.

(6) *How can we fully exploit the clearly evident encoding of identity in prosody and other suprasegmental features of speech?*

Although it is quite obvious that suprasegmental features of speech play important roles in speech perception, it is still difficult to use these features in both speaker recognition and speech recognition. This is because (a) it is difficult to automatically extract these features correctly, (b) we do not know how to model these dynamic features, (c) they are highly variable, and (d) speakers can exert more conscious control over them than over segmental features.

(7) *Is the "sheep and goats" problem (a small percentage of speakers account for the majority of errors) universal?*

Is there any universal set of parameters, algorithms, etc. that is good for all/any set of speakers? Or is the encoding of identity so speaker-dependent that the optimal configuration parameters (algorithm, acoustic features, etc.) depend on the speaker set to be identified? Should we combine separate features to identify a wide range of speakers? Will speaker recognition ever be practical for 100% of the speaking population (or will some speakers need to be excluded because their "identity encoding scheme" cannot be exploited by the system configuration)? Is there any set of parameters that is good for separating speakers whose voices sound identical, such as twins? Can we separate these identically sounding voices? Is it worth trying?

(8) *Can we ever reliably cluster speakers on the basis of similarity / dissimilarity?*

Is there a "universal" (text independent, feature set independent, etc.) relationship between pairs of speakers? Can we make absolute statements about the relationships of speakers (e.g., these two speakers will never be confused)? We can find some systematic speaker variation across phonemes, but we do not know how systematic and how random it is. This question is very closely related to question (7).

(9) *How do we acquire realistically sized (viable) databases that still adequately model inter- and intra-speaker variability?*

How can we model or sample the universal distribution of voices (inter-speaker variability)? How should we choose the speaker set? How should we collect speech data for each speaker in order to capture the intra-speaker variability? How should we select the texts (sentences, words, syllables, etc.)? How should we make common speech databases for evaluating speaker recognition techniques? It is crucial to have common speech databases for comparing the effectiveness of different techniques (Godfrey et al., 1994; Naik et al., 1989). Major speech databases designed for speaker recognition and related areas include the KING corpus and the SWITCHBOARD corpus (Godfrey et al., 1994). It is also important to consider the recording conditions, such as background noise, the type of microphone/telephone, and channel characteristics, in order to collect a

database appropriate for evaluating the robustness of the techniques.

(10) *How do we deal with long-term variability in people's voices?*

How should we update the speaker models to cope with the gradual changes in people's voices? Is there any systematic long-term variation? Can we estimate the session-to-session intra-speaker variability just by using speech samples collected at one session? How do we adequately retrain the models? How can we prevent the models from being updated by the wrong persons' voices?

(11) *Can we model or develop strategies for dealing with factors that significantly alter a person's voice (short-term transitory)?*

Sources of these alterations include short-term illness (flu, etc.), emotion, fatigue, and the words spoken. So far, no one has succeeded in modeling these voice alterations.

(12) *How can we extract text-independent speaker-specific features?*

Various methods have been investigated in text-independent speaker recognition. However, there is still no good method for modeling transitional speaker characteristics. Another important issue is how to extract phoneme-dependent speaker characteristics without knowing the phonemes.

(13) *How can we deal with deliberately disguised voices?*

Are there acoustic features that are invariant regardless of the disguise method? This question is of particular importance in forensic speaker recognition. Criminals may disguise their voices or mimic another person's voice. Does this pose a real threat to speaker recognition systems? How can we cope with this problem?

(14) *How does speaker recognition compare to other means of personal identification, both now and in the future?*

Fingerprints, signatures, and various other means can also be used for identification. The cost/performance ratio of the speaker recognition system must become lower than that of the other means before it can be widely used in the real world. In contrast with finger-printing, we probably have to assume that there are pairs of speakers whose voices cannot be separated when the population is large.

(15) *What are the conditions that speaker recognition systems must satisfy in order for them to be utilized in the field?*

This question is a generalization of question (14). What is speech good for? Cost/performance is not necessarily the most important issue in actual use. Even if the performance is not very high, the system can be used in combination with other means. Although they are difficult to fully understand, the critical conditions that must be met by commercial systems are crucial.

(16) *What about combining speech and speaker recognition?*

The text-prompted method is an example of combining speech and speaker recognition. Since phonetic information and speaker information are correlated, combining the approaches and ideas of speaker and speech recognition is expected to create new ideas and improve recognition performance. But how can we do it? An interesting research topic is the automatic adjustment of speaker-independent phoneme models to each new speaker so that the performance of both speech and speaker recognition are simultaneously improved.

4. Concluding remarks

There have been many recent advances and successes in speaker recognition technology. AT&T and TI (with Sprint) have started field tests and actual application of speaker recognition technology. However, there are still many problems for which good solutions remain to be found. Most of these problems arise from variability, including speaker-generated variability and variability in channel and recording conditions. It is very important to investigate feature parameters that are stable over a long period, insensitive to variations in speaking manner, including speaking rate and level, and robust against variations in voice quality such as those due to voice disguise or colds. It is also important to develop a method to cope with the problems of distortion due to telephone sets and channels and background and channel noises.

Speaker characterization techniques are related to research on improving speech recognition accuracy by speaker adaptation (Furui, 1991a,b), improving synthesized speech quality by adding the natural

characteristics of voice individuality, and converting synthesized voice individuality from one speaker to another. Studies on automatically extracting the speech periods of each person separately from a dialogue involving more than two people have recently appeared as an extension of speaker recognition technology (Gish et al., 1991; Siu et al., 1992; Wilcox et al., 1994). Diversified research related to speaker-specific information in speech waves is expected to increase in the near future.

References

Atal, B., 1972. Automatic speaker recognition based on pitch contours. J. Acoust. Soc. Am. 52 (6), 1687–1697.

Atal, B., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am. 55 (6), 1304–1312.

Carey, M., Parris, E., 1992. Speaker verification using connected words. Proc. Institute of Acoustics 14 (6), 95–100.

Carey, M., Parris, E., Lloyd-Thomas, H., Bennet, S., 1996. Robust prosodic features for speaker identification. In: Proc. Internat. Conf. Spoken Language Processing, Philadelphia, PA, pp. 1800–1803.

Doddington, G., 1985. Speaker recognition-identifying people by their voices. Proc. IEEE 73 (11), 1651–1664.

Eatock, J., Mason, J., 1990. Automatically focusing on good discriminating speech segments in speaker recognition. In: Proc. Internat. Conf. Spoken Language Processing, vol. 5.2, pp. 133–136.

Furui, S., Itakura, F., Saito, S., 1972. Talker recognition by longtime averaged speech spectrum. Trans. IECE A55 1 (10), 549–556.

Furui, S., 1974. An analysis of long-term variation of feature parameters of speech and its application to talker recognition. Trans. IECE A57 (12), 880–887.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. 29 (2), 254–272.

Furui, S., 1986. Research on individuality features in speech waves and automatic speaker recognition techniques. Speech Communication 5 (2), 183–197.

Furui, S., 1989. Digital Speech Processing, Synthesis, and Recognition. Marcel Dekker, New York.

Furui, S., 1991. Speaker-independent and speaker-adaptive recognition techniques. In: Furui, S., Sondhi, M.M. (Eds.), Advances in Speech Signal Processing. Marcel Dekker, New York, pp. 597–622.

Furui, S., 1991b. Speaker-dependent-feature extraction, recognition and processing techniques. Speech Communication 10 (6), 505–520.

Furui, S., 1994. An overview of speaker recognition technology. In: ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1–9.

Gales, M., Young, S., 1993. HMM recognition in noise using parallel model combination. In: Proc. Eurospeech, Berlin, pp. II-837-840.

Gauvain, J., Lamel, L., Prouts, B., 1995. Experiments with speaker verification over the telephone. In: Proc. Eurospeech, Madrid, pp. 651–654.

Gish, H., Siu, M., Rohlicek, R., 1991. Segregation of speakers for speech recognition and speaker identification. In: Proc. IEEE Internat. Conf. on Acoust. Speech, Signal Processing, Toronto, S13.11, pp. 873–876.

Godfrey, J., Graff, D., Martin, A., 1994. Public databases for speaker recognition and verification. In: ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 39–42.

Griffin, C., Matsui, T., Furui, S., 1994. Distance measures for text-independent speaker recognition based on MAR model. In: Proc. IEEE Internat. Conf. on Acoust. Speech, Signal Processing, Adelaide, 23.6, pp. I-309–312.

Higgins, A., Wohlford, R., 1986. A new method of text-independent speaker recognition. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, 17.3, pp. 869–872.

Higgins, A., Bahler, L., Porter, J., 1991. Speaker verification using randomized phrase prompting. Digital Signal Process. 1, 89–106.

Juang, B., Soong, F., 1990. Speaker recognition based on source coding approaches. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, S5.4, pp. 613–616.

Kunzel, H., 1994. Current approaches to forensic speaker recognition. In: ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 135–141.

Li, K., Wrench Jr., E., 1983. An approach to text-independent speaker recognition with short utterances. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, 12.9, pp. 555–558.

Markel, J., Oshika, B., Gray, A., 1977. Long-term feature averaging for speaker recognition. IEEE Trans. Acoust. Speech Signal Process. 25 (4), 330–337.

Markel, J., Davi, S., 1979. Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base. IEEE Trans. Acoust. Speech Signal Process. 27 (1), 74–82.

Martin, F., Shikano, K., Minami, Y., 1993. Recognition of noisy speech by composition of hidden Markov models. In: Proc. Eurospeech, Berlin, pp. II-1031–1034.

Matsui, T., Furui, S., 1990. Text-independent speaker recognition using vocal tract, pitch information. In: Proc. Internat. Conf. on Spoken Language Processing, Kobe, 5.3, pp. 137–140.

Matsui, T., Furui, S., 1991. A text-independent speaker recognition method robust against utterance variations. In: Proc. IEEE Internat. Conf. on Acoust. Speech Signal Processing, S6.3, pp. 377–380.

Matsui, T., Furui, S., 1992. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. In: Proc. IEEE Internat. Conf. on Acoust. Speech, Signal Processing, San Francisco, pp. II-157–160.

Matsui, T., Furui, S., 1993. Concatenated phoneme models for

text-variable speaker recognition. In: Proc. IEEE Internat. Conf. on Acoust. Speech, Signal Processing, Minneapolis, pp. II-391–394.

Matsui, T., Furui, S., 1994. Similarity normalization method for speaker verification based on a posteriori probability. In: ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 59–62.

Matsui, T., Furui, S., 1994. Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker recognition. In: Proc. IEEE Internat. Conf. on Acoust. Speech, Signal Processing, Adelaide, 13.1.

Matsui, T., Furui, S., 1996. Robust methods of updating model and a priori threshold in speaker verification. In: Proc. IEEE Internat. Conf. on Acoust. Speech, Signal Processing, Atlanta, pp. I-97–100.

Matsui, T., Furui, S., 1996b. Speaker recognition using HMM composition in noisy environments. Computer Speech and Language 10, 107–116.

Montacie, C. et al., 1992. Cinematic techniques for speech processing: Temporal decomposition and multivariate linear prediction. In: Proc. IEEE Internat. Conf. on Acoust. Speech, Signal Processing, San Francisco, pp. I-153–156.

Naik, J., Netsch, M., Doddington, G., 1989. Speaker verification over long distance telephone lines. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, S10b.3, pp. 524–527.

Newman, M., Gillick, L., Ito, Y., McAllaster, D., Peskin, B., 1996. Speaker verification through large vocabulary continuous speech recognition. In: Proc. Internat. Conf. Spoken Language Processing, Philadelphia, PA, pp. 2419–2422.

O'Shaugnessy, D., 1986. Speaker recognition. IEEE ASSP Mag. 3 (4), 4–17.

Poritz, A., 1982. Linear predictive hidden Markov models and the speech signal. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, S11.5, pp. 1291–1294.

Reynolds, D., 1994. Speaker identification and verification using Gaussian mixture speaker models. In: ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 27–30.

Rose, R., Reynolds, R., 1990. Text independent speaker identification using automatic acoustic segmentation. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, S51.10, pp. 293–296.

Rose, R., Hofstetter, E., Reynolds, R., 1994. Integrated models of signal and background with application to speaker identification in noise. IEEE Trans. Speech Audio Process. 2 (2), 245–257.

Rosenberg, A., Soong, F., 1987. Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. Computer Speech and Language 22, 143–157.

Rosenberg, A., Lee, C., Soong, F., 1990. Sub-word unit talker verification using hidden Markov models. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing S5.3, pp. 269–272.

Rosenberg, A., Lee, C., Soong, F., McGee, M., 1990. Experiments in automatic talker verification using sub-word unit hidden Markov models. In: Proc. Internat. Conf. on Spoken Language Processing, 5.4, pp. 141–144.

Rosenberg, A., Lee, C., Gokcen, S., 1991. Connected word talker verification using whole word hidden Markov models. In: Proc. IEEE Internat. Conf. on Acoust. Speech, Signal Processing, Toronto, S6.4, pp. 381–384.

Rosenberg, A., Soong, F., 1991. Recent research in automatic speaker recognition. In: Furui, S., Sondhi, M.M. (Eds.), Advances in Speech Signal Processing. Marcel Dekker, New York, pp. 701–737.

Rosenberg, A., 1992. The use of cohort normalized scores for speaker verification. In: Proc. Internat. Conf. on Spoken Language Processing, Banff, Th.sAM.4.2, pp. 599–602.

Savic, M., Gupta, S., 1990. Variable parameter speaker verification system based on hidden Markov modeling. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, S5.7, pp. 281–284.

Setlur, A., Jacobs, T., 1995. Results of a speaker verification service trial using HMM models. In: Proc. EUROSPEECH'95, Madrid, pp. 639–642.

Shikano, K., 1985. Text-independent speaker recognition experiments using codebooks in vector quantization. J. Acoust. Soc. Am. 77, 11, abstract.

Siu, M., Yu, G., Gish, H., 1992. An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers. In: Proc. IEEE Internat. Conf. on Acoust. Speech, Signal Processing, San Francisco, pp. I-189–192.

Soong, F., Rosenberg, A., Juang, B., 1987. A vector quantization approach to speaker recognition. AT&T Tech. J. 66, 14–26.

Soong, F., Rosenberg, A., 1988. On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Trans. Acoust. Speech Signal Process. 36 (6), 871–879.

Sugiyama, M., 1988. Segment based text independent speaker recognition. In: Proc. Spring Meeting of Acoust. Soc. Japan, pp. 75–76 (in Japanese).

Tishby, N., 1991. On the application of mixture AR hidden Markov models to text independent speaker recognition. IEEE Trans. Acoust. Speech, Signal Process. 30 (3), 563–570.

Wilcox, L., Chen, F., Kimber, D., Balasubramanian, V., 1994. Segmentation of speech using speaker identification. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, pp. I-161–164.

Zheng, Y., Yuan, B., 1988. Text-dependent speaker identification using circular hidden Markov models. In: Proc. IEEE Internat. Conf. on Acoust., Speech, Signal Processing, S13.3, pp. 580–582.