

論文 / 著書情報
Article / Book Information

Title	Recent Advances in and Perspectives of Speaker Recognition Technology
著者	古井貞熙
Author	SADAOKI FURUI
Journal/Book name	IWHIT 94, Aizu, , , pp. 35-42
発行日 / Issue date	1994, 9

Recent Advances in and Perspectives of Speaker Recognition Technology

Sadaoki Furui

NTT Human Interface Laboratories

3-9-11, Midori-cho, Musashino-shi, Tokyo, 180 Japan

ABSTRACT

Although many recent advances and successes in speaker recognition have been achieved, there are still many problems for which good solutions remain to be found. This paper first introduces recent advances in the text-independent method, the text-prompted method, and the likelihood normalization technique. Then it describes 21 things we still do not know about speaker recognition. The paper concludes with a short discussion assessing the current status and possibilities for the future.

I. INTRODUCTION

Speaker recognition is the process of automatically recognizing who is speaking by using speaker-specific information included in speech waves [3][4][5][22]. This technique can be used to verify the identity claim of people accessing systems, that is, access control of various services by voice. These 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.

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 as a key to confirm the identity claim of a

speaker are classified as speaker verification.

In 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 the unknown speaker, and an utterance of the unknown speaker is compared with the model for the speaker whose identity is claimed. If the match is high 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 true customers. Conversely, a low threshold enables customers 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.

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 [2][15][21][27]. 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 in which pre-

determined key words cannot be used. Therefore, text-independent methods have recently drawn more attention.

II. RECENT ADVANCES

2.1 Text-Independent Speaker Recognition

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, vector quantization (VQ)-based and ergodic-hidden Markov model (HMM)-based methods are being investigated.

(1) VQ-based methods

VQ codebooks, consisting of a small number of representative feature vectors, are used as an efficient means of characterizing speaker-specific features [8][9][10][20][25][26]. 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.

(2) Ergodic-HMM-based methods

Over a long timescale, the temporal variation in speech signal parameters can be represented by stochastic Markovian transitions between states. Poritz [17] 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 for characterizing the output probability function. He characterized the automatically obtained categories as strong voicing, silence, nasal/liquid, stop burst/post silence, and frication.

Savic et al. [24] also used a five-state ergodic linear predictive HMM for broad phonetic categorization. After identifying frames which belong to particular phonetic categories, 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. Experimental results showed that verification accuracy can be considerably improved by this category-dependent weighted linear combination method.

Matsui et al. [11] compared the VQ-based method with the discrete/continuous ergodic HMM-based method, particularly from the viewpoint of robustness against utterance variations. It was shown 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. It was also shown that the information on transitions between different states is ineffective for text-independent speaker recognition, and, therefore, the speaker recognition rates using the continuous ergodic HMM are strongly correlated with the total number of mixtures, irrespective of the number of states.

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 et al. [18]. Furthermore, the VQ-based method can be viewed 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.

2.2 Text-Prompted Speaker Recognition

Both text-dependent and text-independent methods have a serious problem. That is, they can easily be cheated, because someone can record the voice of a registered speaker uttering key words or sentences and then play it back into a microphone to be accepted as the registered speaker. One way to cope with this problem is to use a small set of words, such as digits, as key words and prompt the user to utter a given sequence of key words that is randomly chosen every time the system is used

[7][21].

Recently, a text-prompted speaker recognition method was proposed in which key sentences are completely changed every time [12][14]. 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 it can also reject utterances whose text differs from the prompted text, even if it is uttered by a registered speaker. A recorded and played back voice can thus be correctly rejected.

Figure 1 shows a block diagram of this method. It 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.

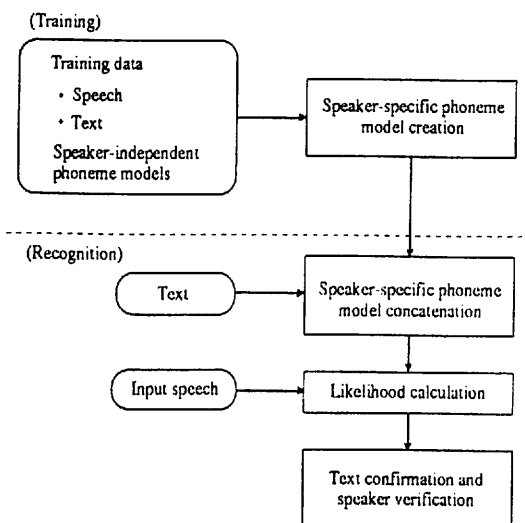


Fig. 1. Block diagram of text-prompted speaker recognition.

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 show that, when the adaptation method for tied-mixture-based phoneme models and the likelihood normalization method were used, a speaker and text verification rate of 99.4% was obtained.

2.3 Likelihood Normalization

The most significant factor affecting automatic speaker recognition performance is variations in signal characteristics from trial to trial (intersession variability or variability over time). Variations are generated by the speaker him/herself, by differences in recording and transmission conditions, and by noise. Speakers cannot repeat an utterance precisely the same way from trial to trial. It is well known that tokens of the same utterance recorded in one session are much more highly correlated than tokens recorded in separate sessions, and there are long term trends in voices.

It is important for speaker recognition systems to accommodate these variations. Two types of normalization techniques have been tried: one in the parameter domain and the other in the distance/similarity domain. A typical normalization technique in the parameter domain is the spectral equalization technique, or "blind equalization" method. It has been shown to be effective in reducing linear channel effects and long-term spectral variation [1][2].

Higgins et al. [7] proposed a normalization method for distance (similarity or likelihood) values that uses a likelihood ratio. The likelihood ratio is the ratio of the conditional probability of the observed measurements of the utterance given the claimed identity to the conditional probability of the observed measurements given the speaker is an impostor. Likelihood ratio normalization approximates optimal scoring in Bayes' sense.

Matsui et al. [12][13] proposed a normalization method based on a posteriori probability. 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. Experimental results indicate that both normalization methods 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 [13][23].

III. WHAT WE DON'T KNOW ABOUT SPEAKER RECOGNITION

(1) How correctly can human beings 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 shown 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. 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 will 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 typical parameters we use are cepstral and delta-cepstral parameters. They 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 transition. We do not know how to represent these transitional features for speaker recognition. As described in 2.1 (2), 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 control over them than over segmental features.

(7) How can we improve the robustness of speaker recognition techniques against noisy speech or speech distorted by telephone?

The robustness issue is crucially important 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 [19].

(8) Is the "sheep and goats" problem (a small percentage of speakers account for a major percentage 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.) are dependent 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 with twins? Can we separate these identically sounding voices? Is it worth trying?

(9) 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 much related to question (8).

(10) How do we capture 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.)?

(11) 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 rates for false acceptance and false rejection are equal. 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. Can we estimate the session-to-session intra-speaker variability by just using speech samples collected at one session?

(12) How can we normalize the intra-speaker variation of likelihood (similarity) values in speaker verification?

To compensate for the intra-speaker variations, normalization techniques using the likelihood ratio or a posteriori probability have been proposed. These methods use a set of speaker models for normalization. How to choose these speakers, or how to make a model covering these speakers is an important issue. In particular how to pick the normalization set of speakers a priori?

(13) 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.

(14) How should we update the speaker models to cope with the gradual changes in people's voices?

How do we adequately solve the retraining problem? How do we deal with long-term variability in people's voices? Is there any systematic long-term variation?

(15) 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. How to extract phoneme-dependent speaker characteristics without knowing the phonemes is also an important issue.

(16) How to prevent speaker verification systems from being defeated by a recorded voice?

To cope with this problem, the text-

prompted speaker recognition method has been proposed. In addition, 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. The text-prompted method is useful in such situations.

(17) 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?

(18) How should we make common speech databases for evaluating speaker recognition techniques?

It is crucially important to have common speech databases for comparing the worth of different techniques [6][16]. This question is very much related to question (10) about the size of databases for modeling inter- and intra-speaker variability. 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.

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

Fingerprints, signatures, and various other means can 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.

(20) 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 (19). What is speech good for? Cost/perfor-

mance 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. It is crucially important, but difficult to fully understand, what are the critical conditions that must be met by commercial systems.

(21) What about combining speech and speaker recognition?

The text-prompted method is an example of combining of 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.

IV. CONCLUSION

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; Sprint's Voice Phone Card is already being used by many customers. However, there are still many problems for which good solutions remain to be found. Speaker characterization techniques are related to research on improving speech recognition accuracy by speaker adaptation, improving synthesized speech quality by adding the natural characteristics of voice individuality, and converting synthesized voice individuality from one speaker to another. I expect that diversified research related to speaker-specific information in speech waves will increase in the near future.

ACKNOWLEDGMENTS

I would like to thank Dr. Michael Barlow, Ms. Tomoko Matsui, and the other members of the Furui Research Laboratory at the NTT Human Interface Laboratories for their valuable comments and suggestions.

REFERENCES

- [1] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," *J. Acoust. Soc. Am.*, Vol. 55, No. 6, pp. 1304-1312 (1974).
- [2] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," *IEEE Trans. Acoust. Speech, Signal Processing*, Vol. 29, No. 2, pp. 254-272 (1981).
- [3] S. Furui, "Speaker-Independent and Speaker-Adaptive Recognition Techniques," in *Advances in Speech Signal Processing* (eds. S. Furui and M. M. Sondhi), Marcel Dekker, New York, pp. 597-622 (1991a).
- [4] S. Furui, "Speaker-Dependent-Feature Extraction, Recognition and Processing Techniques," *Speech Communication*, Vol. 10, No. 5-6, pp. 505-520 (1991b).
- [5] S. Furui, "An Overview of Speaker Recognition Technology," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 1-9 (1994).
- [6] J. Godfrey, D. Graff, and A. Martin, "Public Databases for Speaker Recognition and Verification," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 39-42 (1994).
- [7] A. Higgins, L. Bahler, and J. Porter, "Speaker Verification Using Randomized Phrase Prompting," *Digital Signal Processing*, Vol. 1, pp. 89-106 (1991).
- [8] K. -P. Li and E. H. Wrench Jr., "An Approach to Text-Independent Speaker Recognition with Short Utterances," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 12.9, pp. 555-558 (1983).
- [9] T. Matsui and S. Furui, "Text-Independent Speaker Recognition Using Vocal Tract and Pitch Information," *Proc. Int. Conf. Spoken Language Processing*, 5.3, pp. 137-140 (1990).
- [10] T. Matsui and S. Furui, "A Text-Independent Speaker Recognition Method Robust Against Utterance Variations," *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*, S6.3, pp. 377-380 (1991).
- [11] T. Matsui and S. Furui, "Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMMs," *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, San Francisco, pp. II-157-160 (1992).
- [12] T. Matsui and S. Furui, "Concatenated Phoneme Models for Text-Variable Speaker Recognition," *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, Minneapolis, pp. II-391-394 (1993).
- [13] T. Matsui and S. Furui, "Similarity Normalization Method for Speaker Verification Based on a Posteriori Probability," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 59-62 (1994a).
- [14] T. Matsui and S. Furui, "Speaker Adaptation of Tied-Mixture-Based Phoneme Models for Text-Prompted Speaker Recognition," *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, Adelaide, 13.1 (1994b).
- [15] J. M. Naik, L. P. Netsch, and G. R. Doddington, "Speaker Verification over Long Distance Telephone Lines," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, S10b.3, pp. 524-527 (1989).
- [16] J. Naik, "Speaker Verification over the Telephone Network: Databases, Algorithms and Performance Assessment," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 31-38 (1994).
- [17] A. B. Poritz, "Linear Predictive Hidden Markov Models and the Speech Signal," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, S11.5, pp. 1291-1294 (1982).

- [18] R. C. Rose and R. A. Reynolds, "Text Independent Speaker Identification Using Automatic Acoustic Segmentation," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S51.10, pp. 293-296 (1990).
- [19] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated Models of Signal and Background with Application to Speaker Identification in Noise," IEEE Trans. Speech and Audio Processing, Vol. 2, No. 2, pp. 245-257 (1994).
- [20] A. E. Rosenberg and F. K. Soong, "Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes," Computer Speech and Language, 22, pp. 143-157 (1987).
- [21] A. E. Rosenberg, C. -H. Lee, and S. Gokcen, "Connected Word Talker Verification Using Whole Word Hidden Markov Models," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Toronto, S6.4, pp. 381-384 (1991).
- [22] A. E. Rosenberg and F. K. Soong, "Recent Research in Automatic Speaker Recognition," in *Advances in Speech Signal Processing* (eds. S. Furui and M. M. Sondhi), Marcel Dekker, New York, pp. 701-737 (1991).
- [23] A. E. Rosenberg, "The Use of Cohort Normalized Scores for Speaker Verification," Proc. Int. Conf. Spoken Language Processing, Banff, Th.sAM.4.2, pp. 599-602 (1992).
- [24] M. Savic and S. K. Gupta, "Variable Parameter Speaker Verification System Based on Hidden Markov Modeling," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S5.7, pp. 281-284 (1990).
- [25] K. Shikano, "Text-Independent Speaker Recognition Experiments Using Codebooks in Vector Quantization," J. Acoust. Soc. Am. (abstract), Suppl. 1, No. 77, p. S11 (1985).
- [26] F. K. Soong, A. E. Rosenberg, and B. -H. Juang, "A Vector Quantization Approach to Speaker Recognition," AT&T Technical Journal, No. 66, pp. 14-26 (1987).
- [27] Y. -C. Zheng and B. -Z. Yuan, "Text-Dependent Speaker Identification Using Circular Hidden Markov Models," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S13.3, pp. 580-582 (1988).