論文 / 著書情報
Article / Book Information

Title	Speaker Recognition
Authors	SADAOKI FURUI, Giovanni Battista Varile, Antonio Zampolli
Citation	Survey of the State of the Art in Human Language Technology, , , pp. 36-42
Issue date	1997,
URL	http://www.cambridge.org/

Survey of the State of the Art in Human Language Technology[1]

Editors:
Ronald A. Cole
Joseph Mariani
Hans Uszkoreit
Annie Zaenen
Victor Zue

August 3, 1995

[1]This survey is sponsored jointly by the National Science Foundation and the Directorate XIII-E of the Commission of the European Communities.

- **Domain Adaptation:** How to estimate an effective language model for domains which may not have large online corpora of representative text. Another related problem is topic spotting where the topic-specific language model can be used to model the incoming text from a collection of domain-specific language models.
- **Incorporating Structure:** The current state of the art in language modeling has not been able to improve on performance by the use of the structure (whether surface parse trees or deep structure such as predicate argument structure) that is present in language. A concerted research effort to explore structure-based language models may be the key to significant progress in language modeling. This will become more possible as annotated (parsed) data becomes available. Current research using probabilistic LR grammars, or probabilistic Context-Free grammars (including link grammars) is still in its infancy and would benefit from the increased availability of parsed data.

1.7 Speaker Recognition

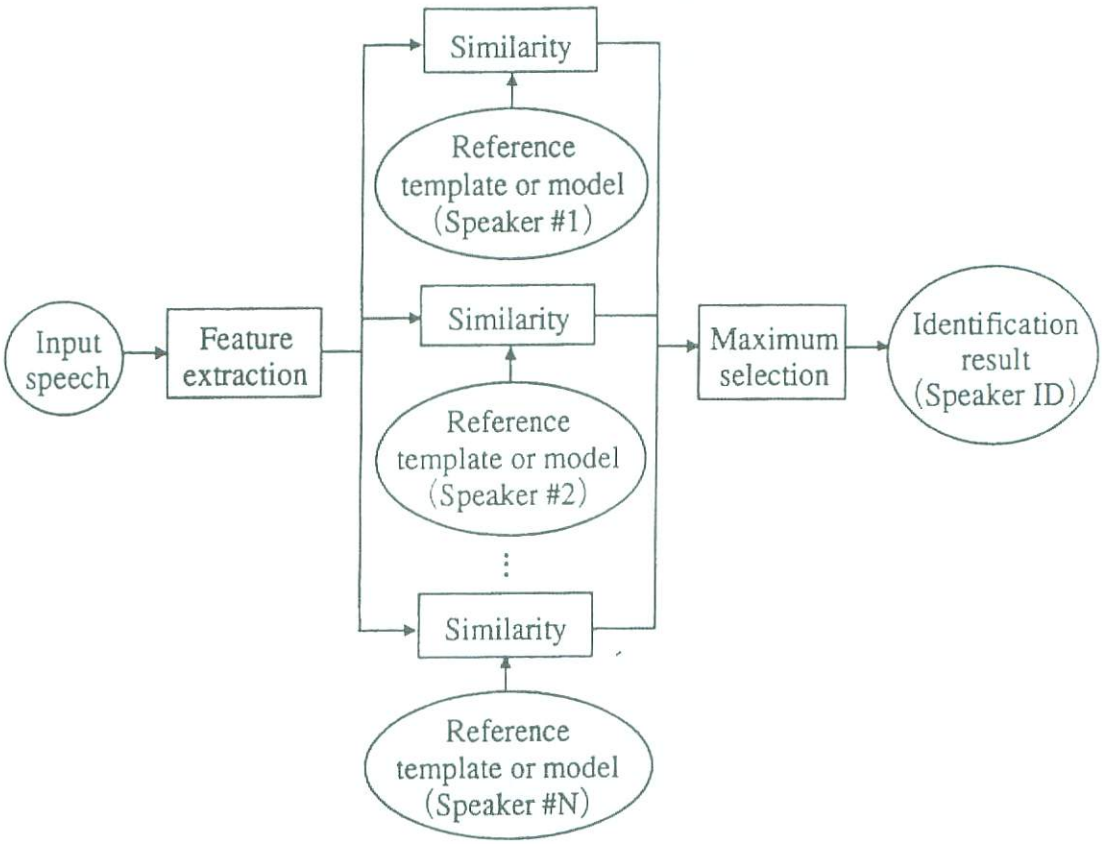
Sadaoki Furui
NTT Human Interface Laboratories, Tokyo, Japan

1.7.1 Principles of Speaker Recognition

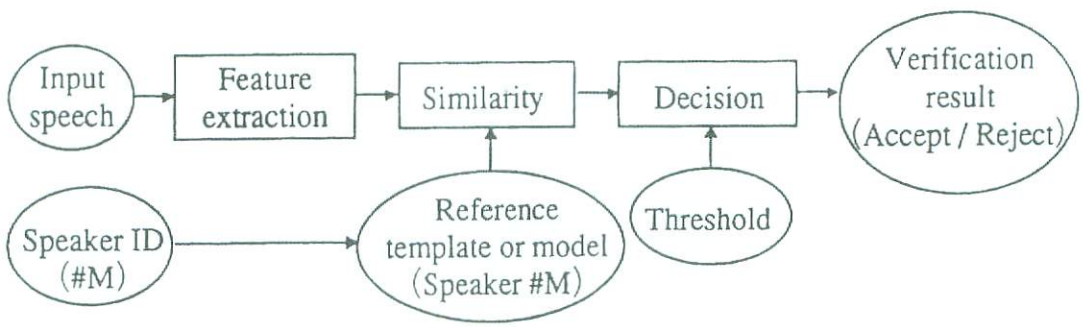
Speaker recognition, which can be classified into identification and verification, is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers. 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. In this way, speaker recognition technology is expected to create new services that will make our daily lives more convenient. Another important application of speaker recognition technology is for forensic purposes.

Figure 1.8 shows the basic structures of speaker identification and verification systems. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. Most applications in which a voice is used as the key to confirm the identity of a speaker are classified as speaker verification.

There is also the case called *open set* identification, in which a reference model for an unknown speaker may not exist. This is usually the case in forensic applications. In this situation, an additional decision alternative, *the unknown*



(a) Speaker identification



(b) Speaker verification

Figure 1.8: Basic structures of speaker recognition systems.

does not match any of the models, is required. In both verification and identification processes, an additional threshold test can be used to determine if the match is close enough to accept the decision or if more speech data needed.

Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former require the speaker to say key words or sentences having the same text for both training and recognition trials, whereas the latter do not rely on a specific text being spoken.

Both text-dependent and independent methods share a problem however. These systems can be easily deceived because someone who plays back the recorded voice of a registered speaker saying the key words or sentences can be accepted as the registered speaker. To cope with this problem, there are methods in which a small set of words, such as digits, are used 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. Yet even this method is not completely reliable, since it can be deceived with advanced electronic recording equipment that can reproduce key words in a requested order. Therefore, a text-prompted (machine-driven-text-dependent) speaker recognition method has recently been proposed by Matsui and Furui (1993b).

1.7.2 Feature Parameters

Speaker identity is correlated with the physiological and behavioral characteristics of the speaker. These characteristics exist both in the spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics and dynamic features spanning several segments).

The most common short-term spectral measurements currently used are Linear Predictive Coding (LPC)-derived cepstral coefficients and their regression coefficients. A spectral envelope reconstructed from a truncated set of cepstral coefficients is much smoother than one reconstructed from LPC coefficients. Therefore it 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 are extracted at every frame period to represent the spectral dynamics. These coefficients are derivatives of the time functions of the cepstral coefficients and are respectively called the delta- and delta-delta-cepstral coefficients.

1.7.3 Normalization Techniques

The most significant factor affecting automatic speaker recognition performance is variation in the signal characteristics from trial to trial (intersession variability and variability over time). Variations arise from the speakers themselves, from differences in recording and transmission conditions, and from background 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 samples recorded in separate sessions. There are also long-term changes 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, the other in the distance/similarity domain.

Parameter-Domain Normalization

Spectral equalization, the so-called *blind equalization* method, is a typical normalization technique in the parameter domain that 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 that use sufficiently long utterances. Cepstral coefficients are averaged over the duration of an entire utterance and the averaged values subtracted from the cepstral coefficients of each frame. Additive variation in the log spectral domain can be compensated for fairly well by this method. However, it unavoidably removes some text-dependent and speaker specific features; therefore it is inappropriate for short utterances in speaker recognition applications.

Distance/Similarity-Domain Normalization

A normalization method for distance (similarity, likelihood) values using a likelihood ratio has been proposed by Higgins, Bahler, et al. (1991). The likelihood ratio is defined as the ratio of two conditional probabilities of the observed measurements of the utterance: the first probability is the likelihood of the acoustic data given the claimed identity of the speaker, and the second is the likelihood given that the speaker is an imposter. The likelihood ratio normalization approximates optimal scoring in the Bayes sense.

A normalization method based on *a posteriori* probability has also been proposed by Matsui and Furui (1994a). The difference between the normalization method based on the likelihood ratio and the method based on *a posteriori* probability is whether or not the claimed speaker is included in the speaker set for normalization; the speaker set used in the method based on the likelihood ratio does not include the claimed speaker, whereas the normalization term for the method based on *a posteriori* probability is calculated by using all the reference speakers, including the claimed speaker.

Experimental results indicate that the two normalization methods are almost equally effective (Matsui & Furui, 1994a). They both improve speaker separability and reduce the need for speaker-dependent or text-dependent thresholding, as compared with scoring using only a model of the claimed speaker.

A new method has recently been proposed in which the normalization term is approximated by the likelihood of a single mixture model representing the parameter distribution for all the reference speakers. An advantage of this method is that the computational cost of calculating the normalization term is very small, and this method has been confirmed to give much better results than either of the above-mentioned normalization methods (Matsui & Furui, 1994a). 1994].

1.7.4 Text-Dependent Speaker Recognition Methods

Text-dependent methods are usually based on template-matching techniques. In this approach, the input utterance is represented by a sequence of feature vectors, generally short-term spectral feature vectors. The time axes of the input utterance and each reference template or reference model of the registered speakers are aligned using a dynamic time warping (DTW) algorithm, and the degree of similarity between them, accumulated from the beginning to the end of the utterance, is calculated.

The hidden Markov model (HMM) can efficiently model statistical variation in spectral features. Therefore, HMM-based methods were introduced as extensions of the DTW-based methods, and have achieved significantly better recognition accuracies (Naik, Netsch, et al., 1989).

1.7.5 Text-Independent Speaker Recognition Methods

One of the most successful text-independent recognition methods is based on vector quantization (VQ). 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. 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 using the codebook of each reference speaker and the VQ distortion accumulated over the entire input utterance is used to make the recognition decision.

Temporal variation in speech signal parameters over the long term can be represented by stochastic Markovian transitions between states. Therefore, methods using an ergodic HMM, where all possible transitions between states are allowed, have been proposed. Speech segments are classified into one of the broad phonetic categories corresponding to the HMM states. After the classification, appropriate features are selected.

In the training phase, reference templates are generated and verification thresholds are computed for each phonetic category. In the verification phase, after the 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 from each category.

This method was extended 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, with additional constraints on the possible transitions between states.

It has been shown that a continuous ergodic HMM method is far superior to a discrete ergodic HMM method and that a continuous ergodic HMM method is as robust as a VQ-based method when enough training data is available. However, when little data are available, the VQ-based method is more robust than a continuous HMM method (Matsui & Furui, 1993a).

A method using statistical dynamic features has recently been proposed. In this method, a multivariate auto-regression (MAR) model is applied to the time series of cepstral vectors and used to characterize speakers. It was reported that identification and verification rates were almost the same as obtained by an HMM-based method (Griffin, Matsui, et al., 1994).

1.7.6 Text-Prompted Speaker Recognition Method

In the text-prompted speaker recognition method, the recognition system prompts each user with a new key sentence every time the system is used and accepts the input utterance only when it decides that it was the registered speaker who repeated the prompted sentence. The sentence can be displayed as characters or spoken by a synthesized voice. Because the vocabulary is unlimited, prospective impostors cannot know in advance what sentence will be requested. Not only can this method accurately recognize speakers, but it can also reject utterances whose text differs from the prompted text, even if it is spoken by the registered speaker. A recorded voice can thus be correctly rejected.

This method is facilitated by using speaker-specific phoneme models as basic acoustic units. One of the major issues in applying this method is how to properly create these speaker-specific phoneme models from 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. In order to properly adapt the models of phonemes that are not included in the training utterances, a new adaptation method based on tied-mixture HMMs was recently proposed by Matsui and Furui (1994b).

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 the input speech matching the sentence model is calculated and used for the speaker recognition decision. If the likelihood is high enough, the speaker is accepted as the claimed speaker.

1.7.7 Future Directions

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. 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 time, insensitive to the variation of speaking manner, including the speaking rate and level, and robust against variations in voice quality due to causes such as voice disguise or colds. It is also important to develop a method to cope with the problem of distortion due to telephone sets and channels, and background and channel noises.

From the human-interface point of view, it is important to consider how the users should be prompted, and how recognition errors should be handled.

Studies on ways to automatically extract 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.

This section was not intended to be a comprehensive review of speaker recognition technology. Rather, it was intended to give an overview of recent advances and the problems which must be solved in the future. The reader is referred to the following papers for more general reviews: Furui, 1986a; Furui, 1989; Furui, 1991; Furui, 1994; O'Shaughnessy, 1986; Rosenberg & Soong, 1991.