# 40 Years of Progress in Automatic Speaker Recognition

Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technology, Japan
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
furui@cs.titech.ac.jp

**Abstract.** Research in automatic speaker recognition has now spanned four decades. This paper surveys the major themes and advances made in the past 40 years of research so as to provide a technological perspective and an appreciation of the fundamental progress that has been accomplished in this important area of speech-based human biometrics. Although many techniques have been developed, many challenges have yet to be overcome before we can achieve the ultimate goal of creating human-like machines. Such a machine needs to be able to deliver satisfactory performance under a broad range of operating conditions. A much greater understanding of the human speech process is still required before automatic speaker recognition systems can approach human performance.

**Keywords:** Speaker recognition, speaker identification, speaker verification, speaker diarization, text-dependent, text-independent, robust recognition.

## 1 Introduction

Speech is the primary means of communication between humans. Speaker recognition is the process of automatically recognizing who is speaking by using the speaker-specific information included in speech waves [16, 17, 18]. Many applications have been considered for speaker recognition. These include secure access control by voice, customizing services or information to individuals by voice, indexing or labeling speakers in recorded conversations or dialogues, surveillance, and criminal and forensic investigations involving recorded voice samples. Currently, the most frequently mentioned application is access control. Access control applications include voice dialing, banking transactions over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, and remote access to computers. Speaker recognition technology, as such, is expected to create new services in smart environments and make our daily lives more convenient.

Recently speaker diarization has been actively investigated, in which an input audio channel is automatically annotated with speakers. Diarization can be used for helping speech recognition, facilitating the searching and indexing of audio archives, and increasing the richness of automatic transcriptions, making them more readable.

This paper reviews major highlights during the last four decades in the research and development of automatic speaker recognition so as to provide a technological perspective. Although considerable technological progress has been made, there still remain many research issues that need to be tackled.

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 1050–1059, 2009.

## 2 40 Years Progress

Topics related to the progress of automatic speaker recognition technology in the past 40 years can be summarized as follows:

### 2.1 1960s and 1970s

(1) ***Early systems*:** The first attempts at automatic speaker recognition were made in the 1960s, one decade later than automatic speech recognition. Pruzansky at Bell Labs [37] was among the first to initiate research by using filter banks and correlating two digital spectrograms for a similarity measure. Pruzansky and Mathews [38] improved upon this technique; and, Li et al. [26] further developed it by using linear discriminators. Doddington at Texas Instruments (TI) [9] replaced filter banks by formant analysis. Intra-speaker variability of features, one of the most serious problems in speaker recognition, was intensively investigated by Endres et al. [11] and Furui [13].

(2) ***Text-independent methods*:** For the purpose of extracting speaker features independent of the phonetic context, various parameters were extracted by averaging over a long enough duration or by extracting statistical or predictive parameters. They include averaged auto-correlation [5], instantaneous spectra covariance matrix [25], spectrum and fundamental frequency histograms [3], linear prediction coefficients [44], and long-term averaged spectra [14].

(3) ***Text-dependent methods*:** Since the performance of text-independent methods was limited, time-domain and text-dependent methods were also investigated [1, 2, 15, 42]. In time-domain methods, with adequate time alignment, one can make precise and reliable comparisons between two utterances of the same text, in similar phonetic environments. Not surprisingly, as a result text-dependent methods were shown to perform significantly better than text-independent methods.

(4) ***Texas Instruments system***: TI built the first fully automated large scale speaker verification system providing high operational security. Verification was based on a four-word, randomized utterance built from a set of 16 monosyllabic words. Digital filter banks were used for spectral analysis, and the decision strategy was sequential, requiring up to 4 utterances for each trial. Several million tests were made over a period of 6 years, for several hundred speakers.

(5) ***Bell Labs system*:** Bell Labs built experimental systems aimed to work over dialed-up telephone lines. Furui [15] proposed using the combination of cepstral coefficients and their first and second polynomial coefficients, now called $\Delta$ and $\Delta\Delta$cepstral coefficients, as frame-based features to increase robustness against distortions by the telephone system. He implemented an online system and tested it for a half year with numerous calls from 120 users. He also proposed methods for updating templates and thresholds for speaker verification decision. The cepstrum-based features later became standard, not only for speaker recognition, but also for speech recognition.

(6) ***Parameter-domain normalization*:** As one typical normalization technique in the parameter domain, spectral equalization, the so-called "blind equalization" method, was confirmed to be effective in reducing linear channel effects and long-term spectral variation. 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 (CMS; cepstral mean subtraction) [2, 15]. This method can compensate fairly well for additive variation in the log spectral domain. This method is especially effective for text-dependent speaker recognition applications using sufficiently long utterances. It has also been shown that Δcepstral coefficients are resistant to linear channel mismatches between training and testing [15].

### 2.2 1980s

(1) ***Statistical modeling*:** Speaker recognition research in the 1980s was characterized by a shift in methodology from the more intuitive template-based approach (a straightforward pattern recognition paradigm) towards a more rigorous statistical modeling framework. Today, most practical speaker recognition systems are based on the statistical framework developed in the 1980s and their results, with significant additional improvements having been made in the 1990s.

(2) ***HMM*:** One of the key technologies developed in the 1980s is the hidden Markov model (HMM) approach [12, 39]. This is a doubly stochastic process in that it has an underlying stochastic process that is not observable (hence the term hidden), but can be observed through another stochastic process that produces a sequence of observations. Although the HMM was well known and understood in a few laboratories (primarily IBM, Institute for Defense Analysis (IDA) and Dragon Systems), it was not until widespread publication of the methods and theory of HMMs in the mid-1980s that the technique became widely applied in virtually every speech recognition research laboratory in the world.

(3) ***HMM-based text-dependent methods*:** As an alternative to the template-matching approach for text-dependent speaker recognition, the HMM technique was introduced in the same way as for speech recognition. HMMs have the same advantages for speaker recognition as they do for speech recognition. Remarkably robust models of speech events can be obtained with only small amounts of specification or information accompanying training utterances. Speaker recognition systems based on an HMM architecture used speaker models derived from a multi-word sentence, a single word, or a phoneme. Typically, multi-word phrases (a string of seven to ten digits, for example) were used, and models for each individual word and for "silence" were combined at a sentence level according to a predefined sentence-level grammar [34].

(4) ***VQ/HMM-based text-independent methods*:** Nonparametric and parametric probability models were investigated for text-independent speaker recognition. As a nonparametric model, vector quantization (VQ) was investigated [43]. A set of short-time training feature vectors of a speaker can be efficiently compressed to a small set of representative points, a so-called VQ codebook. A matrix quantizer encoding multi-frame was also investigated [23, 47]. As a parametric model, the HMM was investigated. Pritz [36] proposed using an ergodic HMM (i.e., all possible transitions between states are allowed). An utterance was characterized as a sequence of transitions through a 5-state HMM in the acoustic feature space. Tishby [48] expanded Poritz's idea by using an 8-state ergodic autoregressive HMM represented by continuous probability density functions with 2 to 8 mixture components per state, which had a higher spectral resolution than the Poritz's

model. Rose et al. [40] proposed using a single-state HMM, which is now called a Gaussian mixture model (GMM), as a robust parametric model.

### 2.3 1990s

(1) ***Robust recognition*:** Research on increasing robustness became a central theme in the 1990s. Matsui et al. [29] 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. They investigated speaker identification rates using the continuous HMM as a function of the number of states and mixtures. It was shown that speaker recognition rates were strongly correlated with the total number of mixtures, irrespective of the number of states. This means that using information about transitions between different states is ineffective for text-independent speaker recognition and, therefore, the GMM achieves almost the same performance as the multiple-state ergodic HMM.

(2) ***Combination of spectral envelope and fundamental frequency features*:** Matsui et al. [28] 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, they used two separate codebooks for voiced and unvoiced speech for each speaker. A new distance measure was introduced to take into account intra- and inter-speaker variability and to deal with the problem of outliers 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. It was 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 cepstral features.

(3) ***HMM adaptation for noisy conditions*:** In order to increase the robustness of speaker recognition techniques against noisy speech, Rose et al. [41] applied the HMM composition (PMC) method [20, 27] to speech recorded 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. In order to cope with the problem of variation in the signal-to-noise ratio (SNR), Matsui et al. [32] 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.

(4) ***Text-prompted method*:** Matsui et al. proposed a text-prompted speaker recognition method, in which key sentences are completely changed every time the system is used [30]. 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 not only accurately recognizes 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.

(5) ***Score normalization*:** How to normalize intra-speaker variation of likelihood (similarity) values is one of the most difficult problems in speaker verification. 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. Likelihood ratio- and a posteriori probability-based techniques were investigated [22, 31]. In order to reduce the computational cost for calculating the normalization term, methods using "cohort speakers" or a "world model" were proposed.

(6) ***Relation to other speech research*:** Speaker characterization techniques were related to research on improving speech recognition accuracy by speaker adaptation [16], improving synthesized speech quality by adding the natural characteristics of voice individuality, and converting synthesized voice individuality from one speaker to another. Studies on speaker diarization, that is, automatically extracting the speech periods of each person separately ("who spoke when") from a dialogue/conversation/meeting involving more than two people appeared as an extension of speaker recognition technology [21, 45, 49]. Speaker segmentation and clustering techniques have been used to aid in the adaptation of speech recognizers and for supplying metadata for audio indexing and searching. This allows for searching audio by speaker and makes speech recognition results easier to read.

### 2.4 2000s

(1) ***Score normalization*:** A family of new normalization techniques has been proposed, in which the scores are normalized by subtracting the mean and then dividing by standard deviation, both terms having been estimated from the (pseudo) imposter score distribution. Different possibilities are available for computing the imposter score distribution: Znorm, Hnorm, Tnorm, Htnorm, Cnorm and Dnorm [4]. State-of-the-art text-independent speaker verification techniques associate one or more parameterization level normalizations (CMS, feature variance normalization, feature warping, etc.) with a world model normalization and one or more score normalizations.

(2) ***Model adaptation*:** Various model adaptation and compensation techniques have been investigated for GMM based speaker recognition methods. McLaren et al. [33] proposed two techniques for GMM mean supervector SVM classifiers: intersession variability modeling and nuisance attribute projection. Petri et al. [35] proposed an unsupervised model adaptation technique including a weighting scheme for the test data, based on the *a posteriori* probability that a test utterance belongs to the target customer model.

(3) ***Combination of audio and visual features*:** There has been a lot of interest in audio-visual speaker verification systems, in which a combination of speech and image information is used. As visual information, lip movement is widely used. The audio-visual combination helps improve system reliability. For instance, while background noise has a detrimental effect on the performance of voice, it

does not have any influence on lip information. Conversely, although the performance of lip recognition systems depends on lighting conditions, lighting does not have any effect on voice quality.

The method of combining two information sources (audio-visual fusion) can be treated as either a classifier combination problem or pattern classification problem. For example, for those systems that can only provide decisions, a majority voting method can be used. If the output of classifiers are compatible (e. g., in the form of posterior probabilities), they can be linearly combined (sum rule) or multiplied together (product or log-linear rule). In addition to these combination methods, researchers have also proposed treating the outputs of individual classifiers as feature vectors and using a classifier such as support vector machines, binary decision trees, and radial basis function networks to classify the vectors [7, 8].

(4) ***High-level features*:** High-level features such as word idiolect, pronunciation, phone usage, prosody, etc. have been successfully used in text-independent speaker verification. Typically, high-level-feature recognition systems produce a sequence of symbols from the acoustic signal and then perform recognition using the frequency and co-occurrence of symbols. In Doddington's idiolect work [10], word unigrams and bigrams from manually transcribed conversations were used to characterize a particular speaker in a traditional target/background likelihood ratio framework. The use of support vector machines for performing the speaker verification task based on phone and word sequences obtained using phone recognizers has been proposed [6]. The benefit of these features was demonstrated in the "NIST extended data" task for speaker verification; with enough conversational data, a recognition system can become "familiar" with a speaker and achieve excellent accuracy. These methods require utterances of at least several minutes long, much longer than those used in conventional speaker recognition methods.

(5) ***MLLR features*:** MLLR (Maximum Likelihood Linear Regression) model adaptation [24] has been widely used in supervised as well as unsupervised HMM adaptation for increasing the robustness of speech recognition. The MLLR has also been widely used in creating text-independent speaker-specific GMMs by adapting speaker-independent GMM (world model). Stolcke [46] proposed using the MLLR adaptation matrix itself as a speaker characterizing feature and reported good experimental results.

## 3 Discussions

### 3.1 Summary of the Technology Progress

In the last 40 years, research in speaker recognition has been intensively carried out worldwide, spurred on by advances in signal processing, algorithms, architectures, and hardware. The technological progress in the last 40 years can be summarized by the following changes [19]:

(1) from template matching to statistical modeling, e.g. HMM and GMM,

(2) from filter bank/spectral resonance to cepstral features (cepstrum + $\Delta$cepstrum + $\Delta\Delta$cepstrum),

(3) from heuristic time-normalization to DTW/DP matching,
(4) from "distance"-based to likelihood-based methods,
(5) from raw scores to normalized scores,
(6) from acoustic features to high-level features,
(7) from maximum likelihood to discriminative approach, e.g. MCE and SVM,
(8) from clean speech to noisy/telephone speech,
(9) from single-modality (audio signal only) to multimodal (audio/visual) recognition,
(10) emergence of application/combination with speech recognition and speech synthesis, e.g. speaker diarization and voice conversion,

Many of these advances have taken place in both the fields of speech recognition and speaker recognition. The majority of technological changes have been directed toward the purpose of increasing robustness of recognition, including many other additional important techniques not noted above.

Although we have witnessed many new technological promises, we have also encountered a number of practical limitations that hinder a widespread deployment of applications and services.

### 3.2 How to Achieve Better Speaker Recognition Performance

There are many outstanding issues and problems in the area of speaker recognition. The most pressing issues, providing challenges for implementing practical and uniformly reliable systems for speaker recognition, are rooted in problems associated with variability and insufficient data. Variability is associated with trial-to-trial variations in recording and transmission conditions and speaking behavior. The most serious variations occur between enrollment sessions and subsequent test sessions, resulting in models that are mismatched to test conditions. Most applications require reliable system operation under a variety of environmental and channel conditions and require that variations in speaking behavior will be tolerated. Insufficient data refers to the unavailability of sufficient amounts of data to train representative models and accurate decision thresholds. Insufficient data is a serious and common problem because most applications require systems that operate with the smallest practicable amounts of training data recorded in the fewest number of enrollment sessions, preferably one.

The challenge is to find techniques that compensate for these deficiencies. A number of techniques have been proposed which provide partial solutions, such as cepstral subtraction techniques for channel normalization and spectral subtraction for noise removal. An especially effective technique for combating both variability and insufficient data is updating models with data extracted from test utterances. Studies have shown that model adaptation, properly implemented, can improve verification performance significantly with a small number of updates. It is difficult, however, for model adaptation to respond to large, precipitous changes. Moreover, adaptation provides for the possibility that customer models might be updated and possibly captured by impostors.

A desirable feature for a practical speaker recognition system is reasonably uniform performance across a population of speakers. Unfortunately, it is typical to observe in a speaker recognition experiment a substantial discrepancy between the best performing individuals, the "sheep", and the worst, the "goats". This additional

problem in variability has been widely observed, but there are virtually no studies focusing on its origin. Speakers with no observable speech pathologies, and for whom apparently good reference models have been obtained, are often observed to be "goats". It is possible that such speakers exhibit large amounts of trial-to-trial variability, beyond the ability of the system to provide adequate compensation.

## 4 Conclusion

Although many important scientific advances have taken place, we have also encountered a number of practical limitations which hinder a widespread deployment of application and services. We still have many research issues as described in the previous section. What we know about human speech processing is very limited. Significant advances in speaker recognition are not likely to come solely from research in statistical pattern recognition and signal processing. Although these areas of investigation are important, the significant advances will come from studies in acoustic-phonetics, speech perception, linguistics, and psychoacoustics.

## References

1. Atal, B.S.: Text-independent speaker recognition: J.A.S.A. 52(181) (A), 83th ASA (1972)
2. Atal, B.S.: Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification: J.A.S.A. 55(6), 1304–1312 (1974)
3. Beek, B., et al.: Automatic speaker recognition system: Rome Air Development Center Report (1971)
4. Bimbot, F.J., et al.: A tutorial on text-independent speaker verification. EURASIP Journ. on Applied Signal Processing, 430–451 (2004)
5. Bricker, P.D., et al.: Statistical techniques for talker identification. B.S.T.J. 50, 1427–1454 (1971)
6. Campbell, W.M., Campbell, J.P., Reynolds, D.A., Jones, D.A., Leek, T.R.: High-level speaker verification with support vector machines. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. I-73–76 (2004)
7. Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer, E., Torres-Carrasquillo, P.A.: Support vector machines for speaker and language recognition. Computer Speech and Language 20(2-3), 210–229 (2006)
8. Cheung, M.-C., Mak, M.-W., Kung, S.-Y.: A two-level fusion approach to multimodal biometric verification. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. V-485-488 (2005)
9. Doddington, G.R.: A method of speaker verification. J.A.S.A. 49(139) (A) (1971)
10. Doddington, G.R.: Speaker recognition based on idiolectal differences between speakers. In: Proc. Eurospeech, pp. 2521–2524 (2001)
11. Endres, W., et al.: Voice spectrograms as a function of age, voice disguise, and voice imitation. J.A.S.A. 49, 6(2), 1842–1848 (1971)
12. Ferguson, J. (ed.): Hidden Markov models for speech, IDA, Princeton, NJ (1980)
13. Furui, S.: An analysis of long-term variation of feature parameters of speech and its application to talker recognition. Electronics and Communications in Japan 57-A, 34–41 (1974)
14. Furui, S., et al.: Talker recognition by long time averaged speech spectrum. Electronics and Communications in Japan 55-A, 54–61 (1972)

15. Furui, S.: Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoustics, Speech, Signal Processing ASSP-29, 254–272 (1981)
16. Furui, S.: Speaker-independent and speaker-adaptive recognition techniques. In: Furui, S., Sondhi, M.M. (eds.) Advances in Speech Signal Processing, pp. 597–622. Marcel Dekker (1991)
17. Furui, S.: Recent advances in speaker recognition. In: Proc. First Int. Conf. Audio- and Video-based Biometric Person Authentication, Crans-Montana, Switzerland, pp. 237–252 (1997)
18. Furui, S.: Digital Speech Processing, Synthesis, and Recognition, 2nd edn. Marcel Dekker, New York (2000)
19. Furui, S.: Fifty years of progress in speech and speaker recognition. In: Proc. 148th ASA Meeting (2004)
20. Gales, M.J.F., Young, S.J.: HMM recognition in noise using parallel model combination. In: Proc. Eurospeech, Berlin, pp. II-837-840 (1993)
21. Gish, H., Siu, M., Rohlicek, R.: Segregation of speakers for speech recognition and speaker identification. In: Proc. ICASSP, S13.11, pp. 873–876 (1991)
22. Higgins, A., et al.: Speaker verification using randomized phrase prompting. Digital Signal Processing 1, 89–106 (1991)
23. Juang, B.-H., Soong, F.K.: Speaker recognition based on source coding approaches. In: Proc. ICASSP, vol. 1, pp. 613–616 (1990)
24. Leggetter, C.J., Woodland, P.C.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language 9, 171–185 (1995)
25. Li, K.P., Hughes, G.W.: Talker differences as they appear in correlation matrices of continuous speech spectra. J.A.S.A. 55, 833–837 (1974)
26. Li, K.P., et al.: Experimental studies in speaker verification using a adaptive system. J.A.S.A. 40, 966–978 (1966)
27. Martin, F., Shikano, K., Minami, Y.: Recognition of noisy speech by composition of hidden Markov models. In: Proc. Eurospeech, Berlin, pp. II-1031–1034 (1993)
28. Matsui, T., Furui, S.: Text-independent speaker recognition using vocal tract and pitch information. In: Proc. Int. Conf. Spoken Language Processing, Kobe, vol. 5.3, pp. 137–140 (1990)
29. Matsui, T., Furui, S.: Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. In: Proc. ICSLP, pp. II-157–160 (1992)
30. Matsui, T., Furui, S.: Concatenated phoneme models for text-variable speaker recognition. In: Proc. ICASSP, pp. II-391–394 (1993)
31. Matsui, T., Furui, S.: Similarity normalization method for speaker verification based on a posteriori probability. In: Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 59–62 (1994)
32. Matsui, T., Furui, S.: Speaker recognition using HMM composition in noisy environments. Computer Speech and Language 10, 107–116 (1996)
33. McLaren, M., Vogt, R., Baker, B., Sridharan, S.: A comparison of session variability compensation techniques for SVM-based speaker recognition. In: Proc. Interspeech, pp. 790–793 (2007)
34. Naik, J.M., et al.: Speaker verification over long distance telephone lines. In: Proc. ICASSP, pp. 524–527 (1989)
35. Petri, A., Bonastre, J.-F., Matrouf, D., Capman, F., Ravera, B.: Confidence measure based unsupervised target model adaptation for speaker verification. In: Proc. Interspeech, pp. 754–757 (2007)

36. Poritz, A.B.: Linear predictive hidden Markov models and the speech signal. In: Proc. ICASSP, vol. 2, pp. 1291–1294 (1982)
37. Pruzansky, S.: Pattern-matching procedure for automatic talker recognition. J.A.S.A. 35, 354–358 (1963)
38. Pruzansky, S., Mathews, M.V.: Talker recognition procedure based on analysis of variance. J.A.S.A. 36, 2041–2047 (1964)
39. Rabiner, L.R., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs (1993)
40. Rose, R., Reynolds, R.A.: Text independent speaker identification using automatic acoustic segmentation. In: Proc. ICASSP, pp. 293–296 (1990)
41. Rose, R.C., Hofstetter, E.M., Reynolds, D.A.: Integrated models of signal and background with application to speaker identification in noise. IEEE Trans. Speech and Audio Processing 2(2), 245–257 (1994)
42. Rosenberg, A.E., Sambur, M.R.: New techniques for automatic speaker verification. IEEE Trans. Acoustics, Speech, Signal Proc. ASSP-23(2), 169–176 (1975)
43. Rosenberg, A.E., Soong, F.K.: Evaluation of a vector quantization talker recognition system in text independent and text dependent models. Computer Speech and Language 22, 143–157 (1987)
44. Sambur, M.R.: Speaker recognition and verification using linear prediction analysis. Ph. D. Dissert., M.I.T (1972)
45. Siu, M., et al.: An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers. In: Proc. ICASSP, pp. I-189–192 (1992)
46. Stolcke, A., Ferrer, L., Kajarekar, S., Shriberg, E., Venkataraman, A.: MLLR transforms as features in speaker recognition. In: Proc. Interspeech 2005, pp. 2425–2428 (2005)
47. Sugiyama, M.: Segment based text independent speaker recognition. In: Proc. Acoust., Spring Meeting of Soc. Japan, pp. 75–76 (1988) (in Japanese)
48. Tishby, N.: On the application of mixture AR hidden Markov models to text independent speaker recognition. IEEE Trans. Acoust., Speech, Signal Processing ASSP-30(3), 563–570 (1991)
49. Wilcox, L., et al.: Segmentation of speech using speaker identification. In: Proc. ICASSP, pp. I-161–164 (1994)