PAMUKKALE ÜNİVERSİTESİ MÜHENDİSLİK FAKÜLTESİ
PAMUKKALE UNIVERSITY ENGINEERING COLLEGE
MÜHENDİSLİK BİLİMLERİ DERGİSİ
JOURNAL OF ENGINEERING SCIENCES
YIL : 2001
CİLT : 7
SAYI : 1
SAYFA : 47-54

FEATURE SELECTION AND CLASSIFICATION TECHNIQUES FOR SPEAKER RECOGNITION

Figen ERTAŞ
Uludağ University, Faculty of Engineering and Architechture, Electronic Engineering Department, Görükle, Bursa

Geliş Tarihi : 22.11.1999

ABSTRACT

Speaker recognition can be considered as a subset of the more general area known as pattern recognition, which may be viewed basically in three stages as: feature selection and extraction, classification, and pattern matching. Extensive research in the past has been directed towards finding effective speech characteristics for speaker recognition. But, so far, no feature set is found to be known to allow perfect discrimination for all conditions. As the performance of features depends on the nature of application, the selection of salient features is a key step in the recognition process. In this paper, we present a general view of speech features and well known classifiers originally developed for text-independent speaker recognition systems. A comparative discussion on choice of suitable speech features and classification techniques is also given.

Key Words : Feature, Classification, LPC, HMM, GMM

KONUŞMACI TANIMA İÇİN ÖZELLİK SEÇİMİ VE SINIFLANDIRMA TEKNİKLERİ

ÖZET

Konuşmacı tanıma; özellik seçip elde etme, sınıflandırma ve örüntü karşılaştırma olarak üç aşamadan oluşan örüntü tanıma olarak bilinen genel bir alanın, bir alt kümesi olarak düşünülebilir. Geçmişten bu yana, konuşmacı tanımaya elverişli ses karakteristiklerinin bulunması yönünde yoğun çalışmalar yapılmış olmasına rağmen, henüz tüm şartlar için mükemmel ayırt etmeye yarayan bir özellik kümesi bulunamamıştır. Dolayısı ile, özelliklerin sistem başarımına etkisi uygulamanın tipine bağlı olduğundan, has özelliklerin seçimi tanıma işleminin en önemli basamağını oluşturmaktadır. Bu makalede, ses özellikleri ve daha çok metinden bağımsız konuşmacı tanıma için geliştirilmiş en çok bilinen sınıflandırma tekniklerine genel bir bakış verilmiştir. Ayrıca, uygun ses özellikleri ve sınıflayıcıların seçimleri karşılaştırmalı olarak tartışılmıştır.

Anahtar Kelimeler : Özellik, Sınıflandırma, Doğrusal öngörülü kodlama, Gizli, Markov modeli, Karma Gaussian modeli

1. INTRODUCTION

This paper presents a general view of firstly the speech features used for generally template based speaker recognition, such as: intensity, formant frequency, pitch, spectral information through filterbank analysis, linear prediction coefficients (LPC), respectively. Then, the most prominent stochastic speaker recognition classifiers originally developed for text-independent (TI) systems, such as: Vector quantisation (VQ), hidden Markov model (HMM), Gaussian mixture model (GMM), and the Neural Networks (NN) are reviewed. For the TI recognition, it is seen that the commonly used features are the cepstral coefficients (Farrell et all., 1994; Mammone et all., 1996). It is also seen that the Gaussian mixture speaker model specifically evaluated for TI speaker identification task using short duration utterances from unconstrained conversational speech provide a robust speaker representation (Reynolds and Rose, 1995).

2. FEATURE PARAMETERS

Many different pieces of information are carried simultaneously in a speech signal. Primarily, they

convey the words that was said and, at a secondary level, they convey information about the identity of the speaker. In addition, speech signals include clues to the physical and emotional state of the speaker, state of the speaker's health, class of the speaker, and the recording environment. Thus, there are large variabilities in the speech signal between speakers and, more importantly, significant variations from instant to instant for the same speaker and text. Broadly speaking, there are two main sources of variation among speakers: anatomical differences and learned differences, which lead to two types of useful features as inherent and learned features. The anatomical differences from speaker to speaker relate to the sizes and shapes of the components of their vocal tracts. For example, a shorter vocal-tract length results in higher formant frequencies, and variations in the size of vocal cords are associated with changes in the average pitch. As a result, inherent features are relatively fixed for a speaker and can be affected by health conditions (e.g., colds that congest the nasal passages). Learned features are not given by nature but are gained through learning to use speakers' speech mechanism and practical use of a language. Learned features might be useful for distinguishing people with similar vocal mechanisms. Such differences reveal themselves in the temporal variations of speech peculiarities of different people. They also affect speaking rate, stress, and melody. As inherent features are less sensitive to counterfeit than learned features, impostors generally find it easier to fool recognizers that are based on learned features than those using inherent features (O'Shaughnessy, 1986).

An important step in the speaker recognition process is to extract sufficient information in some form and size that is amenable to effective modeling. A summary of the features for speaker recognition is reviewed as follows.

2. 1. Intensity

One of the simplest characteristics of any signal is its gain or intensity. The intensity of a speech signal must be defined as a function of time. The source of variations in the intensity of speech are both the subglottal pressure as well as the vocal tract shape as a function of time and represent an important source of speaker-dependent (SD) information in speech. Being relatively easy to measure, a number of systems (particularly earlier ones) have used intensity in conjunction with other parameters (Das and Mohn, 1971). Although it proved an adequate contender in text-dependent (TD) case, the use of the intensity of a speech signal has not been very successful.

2. 2. Formant Frequency

Formant frequencies have always been regarded as suitable candidates among the speech parameters in terms of their suitability for identifying speakers by their voices. But, there are difficulties in their extraction and measurement, especially in the higher formant regions, where much of the SD information is contained (Lewis and Tuthill, 1940). Although formant estimation is a time-consuming process and encounters difficulties with extracting higher order formants, Broad (1972) believes that formants have potential applications in speech and speaker recognition because of their remarkable inter-repetition stability and their close relation to the phonetic concepts of segmentation and equivalence. Nevertheless, they can be used for recognition, with each on its own as well as in combination with other features.

2. 3. Pitch

The frequency of the glottal pulses is one of the important parameters characterizing voiced sounds, corresponding to the fundamental frequency or pitch of the voice. Everyone has a pitch range depending upon their vocal apparatus (e.g. 50-250 Hz for men, 120-500 Hz for women). Many researchers have found pitch as an effective parameter for speaker recognition since it is not sensitive to frequency characteristics of the recording and transmission system while spectral information can be easily affected by the such variations. If the pitch patterns of different speakers are distinct from each other, a recognition system designed on the basis of the pitch would be attainable. When compared to formants, pitch extraction is relatively easier; but it has some disadvantages in terms of their disguisability and inter-repetition stability. Pitch varies significantly in response to factors relating to stress, intonation and emotion. The worst thing about pitch is maybe that it is the easiest acoustic cue which can be disguised. Thus, a system which uses pitch as a parameter may be quite vulnerable to mimics. While the average pitch of a speaker can be easily disguised, it seems unlikely that an impostor could smoothly imitate the whole variation of pitch as a function of time. Atal (1972) found that the pitch contours are strongly SD and yet are quite stable within the utterances of a single speaker. Although, pitch is not good enough to be used alone, it may be used in conjunction with other parameters for speaker recognition, but seldom for word recognition.

2. 4. Filter bank Analysis

Widely used spectral analysis techniques for speech and speaker recognition applications are filter bank

analysis (Davis and Mermelstein, 1980) and LP analysis (Makhoul, 1975). Early TD speaker recognition systems utilized information from the short-time spectrum to provide speaker-specific features. These features consisted of energy measurements from the outputs of a bank of filters. Filter banks, which are a series of adjacent (or overlapping) bandpass filters spanning the useful frequency range of speech, can provide a highly efficient and comprehensive analysis of speech signals. Filter bank processing is often used in combination with other analyses such as pitch, formant, and overall energy. For example, Das and Mohn (1971) used averaged filter bank outputs but supplemented these features with formant data, timing information, and pitch. But these features (spectral patterns) have some disadvantages, because they can be affected by transmission characteristics, and depend upon the level at which the speakers talk and the distance between the speaker and the microphone.

2. 5. LPC Analysis

Short-time spectral information of speech signals is usually extracted through a filter bank, an FFT, or an LPC spectral analysis. The short-term spectrum of the speech signal, defined as a function of time, frequency, and spectral magnitude, is the most common method of representation of the speech signal. Some approaches to the short-term spectrum, such as filter bank magnitudes, LPC spectral and cepstral coefficients, mel-based cepstral coefficients, channel energies in a channel vocoder, or some form of reduced discrete Fourier transform are also popular. They all attempt to capture in a few parameters enough spectral information to identify speakers. Atal (1974) provided a comparison of parameters obtained from LP, the impulse response, autocorrelation, vocal tract area function, and cepstral coefficients, when used with a Mahalanobis distance measure, gave the best speaker recognition performance. Another comparison between the cepstrum and log-area ratios (LAR's) (Furui, 1981) was performed for speaker verification. It was found that cepstral coefficients also outperformed the LAR's. For high quality speech, line spectral pairs (LSP's) were found to yield speaker identification rates that are comparable to or possibly better than those of the cepstral coefficients (Campbell, 1997). However, for telephone quality speech, it was found that cepstral coefficients yield superior performance (Assaleh and Mammone, 1994). As Reynolds and Rose (1995) have also pointed out, LPC cepstral and reflection coefficients have been used extensively for speaker recognition, however, these model-based representations can be strictly affected by noise (Tierney, 1980).

3. CLASSIFICATION TECHNIQUES

A distinguishing feature of a speaker recognition system is whether it is TD or TI. The classification stage is typically to model each speaker, which can be either template or stochastic based generally depending on the dependency to text. The classifier takes the features computed by the feature extractor and performs either template matching or probabilistic likelihood computation on the features, depending on the type of algorithm employed. TD systems naturally use templates, while TI systems use stochastic models. Given the model in TI systems, the pattern matching process is probabilistic and results in a measure of likelihood of the observation to be used by a decision mechanism. But, for template models, the pattern matching is deterministic and produces scores. As we know, theoretically, TI speaker recognition techniques could be used in any situation where TD speaker recognition is applied; the reverse does not hold.

3. 1. Template Based Approach

Prior to the development of probabilistic algorithms, a classical approach to TD speaker recognition is the spectral template matching or spectrogram approach. In this approach, each speaker is represented by a sequence of feature vectors, generally short-term spectral feature vectors, analyzed for each word or phrase. When two persons speak the same utterance, their articulation is similar but not identical; thus, spectrograms of these utterances will be similar but not just the same. Even when the same speaker utters the same word on two different occasions, there are also similarities and differences. Variations may arise from differences in recording, transmission conditions, and voice. But the most significant is the variation produced by the same speaker, which can be voluntary or involuntary. These variations may become so large as to render any speaker recognition decision completely unreliable. Even under the same conditions, speakers cannot repeat an utterance precisely the same way from trial to trial. An important ingredient of a TD speaker recognition system is a means for normalizing trial-to-trial timing variations of utterances of the same phrase by the same speaker. Moderate differences in the timing of speech events can be normalized by aligning the analyzed feature vector sequence of a test utterance to the template feature vector sequence using a dynamic time warp (DTW) algorithm.

Test utterances are compared to training templates by the distance between feature means. All variations to the technique arise from the choices of

feature vectors and distance metrics. Several metrics can be used for minimum distance classifiers of which the Euclidean is the best known and one of the easiest to compute. Later, it was shown that Mahalanobis and weighted distances could further improve discrimination (Ren-hua et all., 1990).

3. 2. Stochastic Based Approach

In a TI mode, the words or phrases used in recognition trials cannot generally be predicted. Therefore, it is not possible to model or match speech events at the level of word or phrases. Probabilistic modeling of speakers refers to modeling speakers by probability distributions rather than by average features and to basing classification decisions on probabilities or likelihoods rather than distances to average features. The following four kind of methods have been discussed for TI speaker recognition.

3. 2. 1. Vector Quantization (VQ)

A set of short-term training feature vectors of a speaker can be used directly to represent the essential characteristics of that speaker. But such a direct representation is impractical when the number of training vectors is large, as the memory and amount of computation required become prohibitively large. Therefore, efficient ways of compressing the training data have been tried using vector quantization (VQ) techniques. The VQ-based method using speaker-specific codebooks, as illustrated in Figure 1, is one of the well-known methods which has been successfully applied to speaker recognition. 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 (Soong et all., 1985; Rosenberg and Soong, 1987; Matsui and Furui, 1991). The key issues in implementing VQ concern the design and search of the codebook, whose creation involves the analysis of a large training sequence of speech (a few minutes long) that sufficiently contains examples of phonemes in many different contexts.

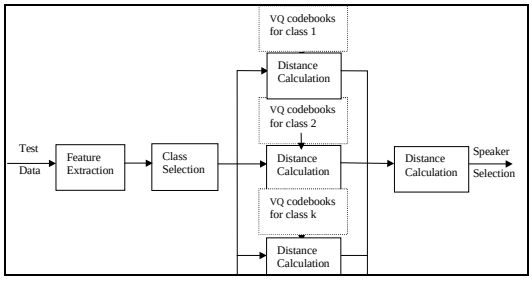

Figure 1. Block diagram of speaker identification system for vector quantizer (VQ) classifier

A VQ codebook is designed by standard clustering procedures for each enrolled speaker using his training data, usually based upon reading a specific text. Since the clustering procedure used to form the codebook averages out temporal information from the code words, there is no need to perform a time alignment. This greatly simplifies the system. In their simplest form, codebooks have no explicit time information, either in terms of temporal order or relative durations, since the codebook entries are not ordered and can derive from any part of the training words. However, implicit durational cues are partially preserved because the entries are chosen to minimize the average distance across all training frames, and frames corresponding to longer acoustic segments (i.e., vowels) are more frequent in the training data. Such segments are more likely to specify codeword positions than the less frequent consonant frames, especially in small codebooks. Besides avoiding segmentation and allowing short test utterances, VQ is computationally efficient as compared to storing and comparing large amounts of template data in the form of individual spectra. Thus, VQ can also be useful for TD, as well as TI recognition.

This method is robust against utterance variations, such as session-to-session variation and TD variation, if sufficient training and test data are available (Soong et all., 1985). However, Matsui and Furui (1991) have reported a VQ-based method that is robust against utterance variations even when only a short utterance is available.

3. 2. 2. Hidden Markov Modeling (HMM)

A possible way to incorporate temporal correlations in the VQ model is by a Markov source of information, or a hidden Markov model (Savic and Gupta, 1990; Tishby, 1991). In conventional Markov models, as illustrated in Figure 2, the speech signal is considered as a sequence of Markov states representing transitions from one speech events to another. The Markov states themselves are “hidden” but are indirectly observable from the sequence of spectral feature vectors. The hidden Markov model

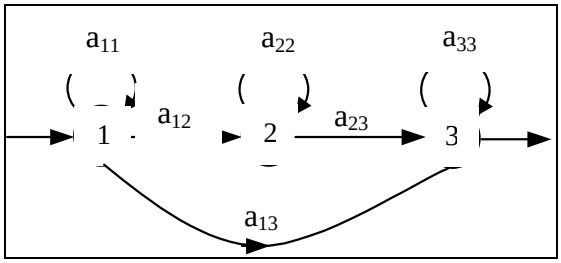

Figure 2. An example of a three-state HMM

(HMM) can only be viewed through another set of stochastic processes that produce the sequence of observations. On a long-time scale, temporal variation in speech signal parameters can be represented by stochastic Markovian transitions between states. The parameters of a HMM are the transition probabilities of observing spectral vectors in each state. Speech events, e.g., words, subwords phonelike units, or acoustic segment units, are represented not just by characteristics sets or sequences of feature vectors but also by probabilistic models of these events calculated from feature "observations" in the training utterances with the HMM representation.

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 categories corresponding to the HMM states. Savic and Gupta (1990) also used a five-state ergodic linear predictive HMM for broad phonetic categorization. If the signal in each state is modeled by an autoregressive source, a special type of HMM results, which is called AR or linear predictive HMM (Poritz, 1982). Autoregressive HMMs, when trained properly, can be used for statistically characterizing speakers, in a TI manner. 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.

HMM-based methods have been shown to be comparable in performance to conventional VQ methods in TI testing (Tishby, 1991) and more recently to outperform conventional methods in TD testing (Reynolds and Carlson, 1995). HMMs in a variety of forms, have been used as probabilistic speaker models for both TI and TD speaker recognition (Poritz, 1982; Matsui and Furui, 1994).

3. 2. 3. Gaussian Mixture Model (GMM)

Rose and Reynolds (1990) investigated a technique based on maximum likelihood estimation of a Gaussian mixture model representation of speaker identity, as illustrated in Figure 3. This method corresponds to the single-state continuous ergodic HMM investigated by Matsui and Furui (1994). Gaussian mixture models were motivated for modeling speaker identity based on two interpretations. First, the individual component Gaussians in a speaker dependent GMM are interpreted to represent some broad phonetic events,

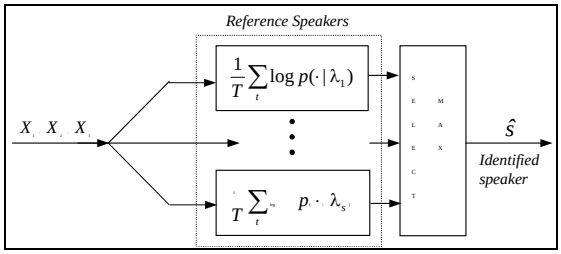

Figure 3. A block diagram of the speaker identification system for Gaussian mixture model

such as vowels, nasals, or fricatives. These acoustic classes reflect some general SD vocal tract configurations that are useful for characterizing speaker identity. By modeling the underlying acoustic classes, the speaker model is better able to represent the short-term variations of a person's voice, allowing high identification performance for short utterances. The second motivation for using Gaussian mixture densities for speaker identification is the empirical observation that a linear combination of Gaussian basis functions is capable of representing a large class of sample distributions.

One of the powerful attributes of the GMM is its ability to form smooth approximations to arbitrarily shaped densities. Reynolds and Rose (1995) reported that the results indicate that GMMs provide a robust speaker representation for the difficult task of speaker identification using corrupted, unconstrained speech. The models are computationally inexpensive and easily implemented on a real-time platform (Reynolds, 1992). Furthermore, their probabilistic framework allows direct integration with speech recognition systems (Reynolds and Heck, 1991) and incorporation of newly developed speech robustness techniques (Rose et all., 1994).

3. 2. 4. Neural Networks (NN)

Several different networks such as multilayer perceptrons (MLP's), for which an example is shown in Figure 4, (Oglesby and Mason, 1990; Rudasi and Zahorian, 1991), time delay neural

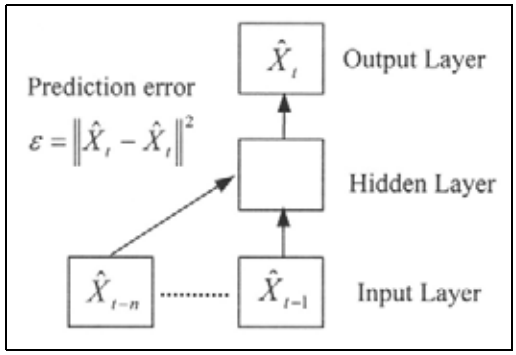

Figure 4. Structure of predictive neural network

networks (TDNN's) (Bennani and Gallinari, 1991), learning vector quantization (LVQ) (Bennani and Gallinari, 1994), have also been applied to various speaker recognition tasks. Rather than training individual models to represent particular speakers, discriminative NN's are trained to model the decision function which best discriminates speakers within a known set. However, readers are referred to Bennani and Gallinari (1994) for a brief review of the current research on neural network systems for speaker recognition.

The Radial Basis Function (RBF) networks (Oglesby and Mason, 1991) share the same underlying structure as the GMM, but model the feature space in different ways. The RBF differs from the above models in that it focuses on modeling the boundary regions separating speaker distributions in the feature space. It uses a pool of basis functions to represent all speakers. However, the basis functions are fixed during training and the speaker's connection weights are trained using a discriminative criterion.

4. DISCUSSION

Early works on speaker identification were dominated by template models. They mostly used filterbanks for frequency analysis and did not intensively exploit the temporal dimension of the information in the speech signals. The trend of research in the last two decades has shown that much of the effort were put into stochastic or its combination with spectral methods, such as Cepstrum, LPC, HMM, VQ, NN, RBF, and more recently GMM.

Among them, VQ neglects speaker-dependent temporal information that may be present in the context, but HMMs make use of the underlying speech sounds as well as the temporal sequencing among these sounds. The GMM can also use temporal information, i.e., the state transition probabilities, in which case a continuous density HMM (Rosenberg et all., 1991) results. However, Tishby (1991) claims that the improvements due to the addition of temporal information are negligible for TI speaker recognition, since the sequencing of temporal information found in the training data does not necessarily reflect the same sequence found in the testing data. Neural networks (NNs) generally require a smaller number of parameters than independent speaker models and have produced good speaker recognition performance, comparable to that of VQ systems. But, the major drawback to many of the NN techniques is that the complete network must be retrained when a new speaker is added to the system (Reynolds and Rose, 1995).

A summary of the most recent techniques for speaker recognition task (identificatin/verification) that appeared in the literature is given in Table 1. The table contains the features and the speech material used (such as sentences, words or digits), percentage of correct identification (or error rates), contribution and the result of each study.

5. REFERENCES

Assaleh, K. T. and Mammone, R. J. 1994. New LP-Derived Features for Speaker Identification. IEEE SAP, Vol. SAP-2, No. 4, 630-638.

Atal, B. S. 1972. Automatic Speaker Recognition Based on Pitch Contours. JASA, 52, 1687-1697.

Atal, B. S. 1974. Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification. JASA, 55, 6, 1304-1312.

Bennani, Y. and Gallinari, P. 1991. On the use of TDNN-Extracted Features Information in Talker Identification. IEEE Proc. ICASSP"91. 265-268.

Bennani, Y. and Gallinari, P. 1994. Connectionist Approaches for Automatic Speaker Recognition. Proc. ESCA Workshop. 95-102.

Broad, D. J. 1972. Formants in Automatic Speech Recognition. Int. J. Man-Machine Studies, 4, 411.

Campbell, J.P. 1997. Speaker Recognition: A Tutorial. Proc. IEEE, Vol. 85, No. 9, Sept. 1997, 1437-1462.

Das, S. K. and Mohn, W. S. 1971. A Scheme for Speech Processing in Automatic Speaker Verification. IEEE Trans. Audio and Electroacoustics, AU-19, 32-43.

Davis, S. B. and Mermelstein, P. 1980. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE ASSP-28. 357-366.

Farrell, K. R. Mammone, R. J. and Assaleh, K. T. 1994. Speaker Recognition Using Neural Networks and Conventional Classifiers. IEEE SAP, Vol. SAP-2, No. 1, Pt. 2, 194-205.

Furui, S. 1981. Cepstral Analysis Technique for Automatic Speaker Verification. IEEE ASSP-29, No. 2, 254-272.

Gopalan, K., Anderson, T.R. and Cupples, E. J. 1999. A Comparison of Speaker Identification Results Using Features Based on Cepstrum and Fourier-Bessel Expansion. IEEE SAP, Vol. SAP-7, No. 3, 289-294.

Table 1. Summary of Speaker Identification (SI) and Speaker Verification (SV) Studies

Haydar, A., Demirekler, M. and Yurtseven, M. K. 1998. Speaker Identification Through Use of Features Selected Using Genetic Algorithm. Electronics Letters, 34 (1), 39-40.

Le Floch, J. L., Montacie, C. and Caraty, M. J. 1995. Speaker Recognition Experiments on the TIMIT Database. Proc. ESCA Workshop. 379-382.

Lewis, D. and Tuthill, C. 1940. Resonant Frequencies and Damping Constants of Resonators Involved in the Production of Sustained Vowel 'O' and 'Ah'. JASA, 11, 451.

Liu, C. S., Wang, H. C. and Lee, C. H. 1996. Speaker Verification Using Normalized Log-likelihood Score. IEEE SAP, Vol. SAP-4, No. 1, 56-59.

Makhoul, J. 1975. Linear Prediction: A Tutorial Review. Proc. IEEE, Vol. 63, 561-580.

Mammone, R. J., Zhang, X. and Ramachandran, R. 1996. Robust Speaker Recognition: A Feature-Based Approach. IEEE Signal Proc. Mag. 58-71.

Matsui, T. and Furui, S. 1991. A Text-Independent Speaker Recognition Method Robust Against Utterance Variations. IEEE Proc. ICASSP'91. 377-380.

Matsui, T. and Furui, S. 1994. Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMM's. IEEE SAP, Vol. SAP-2, No. 3, 456-458.

Murthy, H. A., Beaufays, F., Heck, L.P. and Weintraub, M. 1999. Robust Text-Independent Speaker Identification Over Telephone Channels. IEEE SAP, Vol. SAP-7, No. 5, 554-568.

Oglesby, J. and Mason, J. S. 1990. Optimization and Neural Models for Speaker Identification. IEEE Proc. ICASSP'90. 261-264.

Oglesby, J. and Mason, J. S. 1991. Radial Basis Function Networks for Speaker Recognition. Proc. ESCA Workshop. 87-90.

O'Shaughnessy, D. 1986. Speaker Recognition. IEEE ASSP Magazine. 4-17.

Poritz, A.B. 1982. Linear Predictive hidden Markov Models and the Speech Signal. IEEE Proc. ICASSP"82. 1291-1294.

Ren-hua, W., Lin-shen, H. and Fujisaki, H. 1990. A Weighted Distance Measure Based on the Fine Structure of Feature Space: Application to Speaker Recognition. IEEE Proc. ICASSP'90. 273-276.

Reynolds, D. A. and Heck, L. P. 1991. Integration of Speaker and Speech Recognition Systems. IEEE Proc. ICASSP"91. 869-872.

Reynolds, D. A. 1992. A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification. PhD. Thesis. Georgia Ins. of Tech.

Reynolds, D. A. and Carlson, B. 1995. Text-Independent Speaker Verification Using Decoupled and Integrated Speaker and Speech Recognizers. Proc. EUROSPEECH. 647-650.

Reynolds, D. A. and Rose, R.C. 1995. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE SAP-3, No.1, 72-83.

Rose, R. C. and Reynolds, D. A. 1990. Text-Independent Speaker Identification Using Automatic Acoustic Segmentation. IEEE Proc. ICASSP'90. 293-296.

Rose, R. C., Hofstetter, E.M. and Reynolds, D. A. 1994. Integrated Models of Speech and Background With Application to Speaker Identification in Noise. IEEE SAP-2. No. 2. 245-257.

Rosenberg, A. E. and Soong, F. K. 1987. Evaluation of a Vector Quantization Talker Recognition System in Text-independent and text-dependent modes. Computer Speech and Language. 143-157.

Rosenberg, A.E., Lee, C.-H. and Gokcen, S. 1991. Connected Word Talker Verification Using Whole Word HMMs. IEEE Proc. ICASSP'91, 381-384.

Rudasi, L. and Zahorian, S. A. 1991. Text-Independent Talker Identification With Neural Networks. IEEE Proc. ICASSP"91. 389-392.

Savic, M. and Gupta, S. K. 1990. Variable Parameter Speaker Verification System Based on Hidden Markov Modeling. IEEE Proc. ICASSP"90. 281-284.

Soong, F. K., Rosenberg, A.E., Rabiner, L. R. and Juang, B. H. 1985. A Vector Quantization Approach to Speaker Recognition. IEEE Proc. ICASSP'85, Tamoa, FL, 387-390.

Sukkar, R. A., Gandhi, M. B. and Setlur, A. R. 2000. Speaker Verification Using Mixture Decomposition discrimination. IEEE SAP, Vol. SAP-8, No. 3, 292-299.

Tierney, J. 1980. A study of LPC Analysis of Speech in Additive Noise. IEEE ASSP-28. 389-397.

Tishby, N. Z. 1991. On the Application of Mixture AR Hidden Markov Models to Text-Independent Speaker Recognition. IEEE ASSP-39. 563-570.

Yuan, Z. X., Xu, B. L. and Yu, C. Z. 1999. Binary Quantization of Feature Vectors for Robust Text-Independent Speaker Identification. IEEE SAP, Vol. SAP-6, No. 1, 70-78.

Zhang, Y., Zhu, X. and Zhang, D. 1999. Speaker Verification by Removing Common Information. Electronics Letters, 35 (23), 2009-2011.