IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

# SPEAKER VERIFICATION A QUICK OVERVIEW

Hervé Bourlard [1,2,3] Nelson Morgan [3,4]

IDIAP–RR 98-12

AUGUST 1998

TO APPEAR IN
(1) Traitement de la Parole (Presses Polytechniques et Universitaires Romandes, Lausanne), by R. Boite, H. Bourlard, T. Dutoit, H. Leich, J. Hancq, and (2) Audio Signal Processing in Humans and Machines (publisher still to be defined), by Ben Gold and Nelson Morgan.

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone $+41-27-721\ 77\ 11$
fax $+41-27-721\ 77\ 12$
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[1] Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), Martigny, Switzerland
[2] Swiss Federal Institute of Technology, Lausanne, Switzerland
[3] International Computer Science Institute, Berkeley, CA
[4] University of California, Berkeley, CA

IDIAP Research Report 98-12

# SPEAKER VERIFICATION
# A QUICK OVERVIEW

Hervé Bourlard Nelson Morgan

AUGUST 1998

TO APPEAR IN
(1) Traitement de la Parole (Presses Polytechniques et Universitaires Romandes, Lausanne), by R. Boite, H. Bourlard, T. Dutoit, H. Leich, J. Hancq, and (2) Audio Signal Processing in Humans and Machines (publisher still to be defined), by Ben Gold and Nelson Morgan.

# 1 Introduction

Speech contains many characteristics that are specific to each individual, many of which are independent of the linguistic message for an utterance, and which, in speech recognition, are generally considered as a source of degradation. For instance, each utterance from an individual is produced by the same vocal tract, tends to have a typical pitch range (particularly for each gender), and has a characteristic articulator movement that is associated with dialect or gender. All of these factors have a strong effect on the speech that is highly correlated with the particular individual who is speaking. For this reason, listeners are often able to recognize the speaker identity fairly quickly, even over the telephone. Artificial systems recognizing speakers rather than speech have been the subject of much research over the last 20 years, and commercial systems are already in use.

**Speaker recognition** is a generic term for the classification of a speaker's identity from an acoustic signal. In the case of **speaker identification**, the speaker is classified as being one of a finite set of speakers. As in the case of speech recognition, this will require the comparison of a speech utterance with a set of references for each potential speaker. For the case of **speaker verification**, the speaker is classified as having the purported identity or not. That is, the goal is to automatically accept or reject an identity that is claimed by the speaker. In this case, the user will first identify herself/himself (e.g., by introducing or uttering a PIN code), and the distance between the associated reference and the pronounced utterance will be compared to a threshold that is determined during training. Speaker recognition can be based on text-dependent or text-independent utterances, depending on whether or not the recognition process is constrained to a pre-defined text or not.

Speaker recognition has many potential applications, including: secured use of access cards (e.g., calling and credit credits), access control to databases (e.g., telephone and banking applications), access control to facilities, electronic commerce, information and reservation services, remote access to computer networks, etc.

Speaker identification and verification each require the calculation of a score reflecting the distance between an utterance and a set of references. One of the simplest approaches that was initially used was representing each speaker by a single Gaussian (or a set of Gaussians) in the acoustic parameter space. The parameters were estimated on a training set containing several sentences pronounced by each of the potential user. Assuming that all speakers were equiprobable and that the cost associated with an error is the same for every speaker, the decision rule consisted in assigning an utterance to the speaker with the closest density function. Variants of this approach are still used today and will be briefly discussed in Section 5.

As for the case of speech recognition, speaker recognition has benefited from extensive applications of Hidden Markov Model (HMM) technology in the 1990's. The resulting approaches for these two application areas are very similar. In speaker recognition, though, each speaker is represented by one or several specific HMMs. In the case of text-independent speaker recognition, these HMM models will often be ergodic (fully connected). For text-dependent speaker recognition, specific sentences can be used to model the lexical information in addition to the speaker characteristics. During speaker identification, these models will be used to compute matching scores associated with the input utterance, and the matching speaker will be the one associated with the closest reference. In the case of speaker verification, this matching score will be computed for the model associated with the claimed identity and some number of alternate models; the putative speaker model will then either be accepted or rejected based on some measure of its distance from the models of its rivals (e.g., from its closest rival). The decision also incorporates a threshold that is determined during the enrollment of each new speaker.

For a good introduction to speaker recognition, we refer the reader to [4, 5].

# 2 Acoustic parameters

In speech recognition, the main goal of the acoustic processing module is to extract features that are invariant to the speaker and channel characteristics, and are representative of the lexical content. On the other hand, speaker recognition requires the extraction of speaker characteristic features, which may be independent of the particular words that were spoken. Such characteristics include the gross properties of the spectral envelope (such as the average formant positions over many vowels) or the average range of fundamental frequency. Unfortunately, since these features are often difficult to estimate reliably, particularly for a short enrollment period, current systems often use acoustic parameters that have been developed for use in speech recognition. However, LPC parameters (or LPC cepstra), which have fallen out of favor in automatic speech recognition (ASR) due to their strong dependence on individual speaker characteristics, tend to be preferred in speaker recognition for this very reason. In general, though, features that are based on some kind of short-term spectral estimate are used in speaker recognition much as they are in ASR. In addition, though, pitch information is sometimes used if it can be estimated reliably [1, 10].

Finally, the effects of transmission channel variability are usually reduced by using techniques initially proposed for speech recognition, such as cepstral mean subtraction or RASTA-PLP. However, these techniques also can filter out important speaker-specific characteristics. Further research thus seems to be necessary here.

## 3 Similarity measures

As we have seen for speech recognition, the problem of speaker recognition can be formulated in terms of statistical pattern classification, and the probability that a speaker $S_c$ (rather than some other speaker) has pronounced the sentence associated with the acoustic parameter sequence $X$ is given by:

$$P(S_c|X) \;=\; \frac{P(X|S_c)P(S_c)}{\sum_{i=1}^{I} P(X|S_i)P(S_i)} \tag{1}$$

where $S_c$ represents the identity being tested (or claimed, in the case of speaker verification), and $P(X|S_i)$ the conditional probability of sequence $X$ given the speaker $S_i$. Ideally, the sum in the denominator should include all possible speakers. In general, this sum will be very large, and in the case of speaker verification, should include all possible rival speakers, which is unfortunately impossible.

As for speech recognition, parameters for density estimation in speaker recognition are determined during a training phase (based on maximum likelihood or discriminant criterion).

Once these parameters are determined, the denominator in (1) is independent of the class and can be neglected for speaker identification. Consequently, and assuming equal prior probabilities for all speakers, a speaker $S$ will be identified as speaker $S_c$ if:

$$P(X|S_c) \geq P(X|S_i),\ \forall i \neq c \tag{2}$$

Speaker verification, however, is a form of hypothesis test. In this case, we will verify the hypothesis that a speaker $S$ is indeed the putative speaker $S_c$ if:

$$P(S_c|X) > P(\overline{S_c}|X) \tag{3}$$

where $\overline{S_c}$ represents the set of all possible rival speakers, and the right hand side is the probability of the speaker being anyone except $S_c$. Typically, this is stated with some (speaker dependent) margin or threshold, i.e., a speaker $S$ is taken to be the speaker $S_c$ if:

$$\frac{P(S_c|X)}{P(\overline{S_c}|X)} > \delta_c \tag{4}$$

where $\delta_c$ is a threshold $> 1$, which will have to be optimized independently for each potential customer $S_c$ to guarantee optimal performance of the system (see Section 7). Recall that

$$P(\overline{S_c}|X) = P(S_1 \text{ or } S_2 \text{ or } \ldots \text{ or } S_{i\neq c}|X) = \sum_{i\neq c} P(S_i|X)$$

if events $S_i$ are independent (which is the case) and collectively exhaustive (which will often be wrong). Using (1), and assuming uniform priors $P(S_i)$ over all speakers, criterion (3) becomes:

$$S = S_c \quad \text{if} \quad \frac{P(X|S_c)}{P(X|\overline{S_c})} = \frac{P(X|S_c)}{\sum_{i\neq c} P(X|S_i)} > \delta_c \tag{5}$$

defined as the **likelihood ratio criterion**, and where the sum over $i$ incorporates all the possible speakers.[1] Based on the logarithm of the likelihood ratio, we then have:

$$S = S_c \quad \text{if} \qquad \log P(X|S_c) - \log P(X|\overline{S_c}) > \Delta_c \tag{6}$$

where $\Delta_c = \log \delta_c$. In general, a large enough value of this difference will mean that the identity of $S_c$ is validated, while it will be rejected in the case of a value below the threshold.

Finally, the design of a speaker verification system will thus involve:

1. The optimal setting of the **decision threshold**. As discussed in Section 7, this threshold will usually be estimated by assuming that the distributions of $P(X|S_c)$ and $P(X|\overline{S_c})$ are Gaussian. Figure 1 represents the typical Gaussian approximation to the distributions of likelihoods $P(X|S_c)$ and $P(X|\overline{S_c})$ for a specific training and test set. The variability of these distributions also shows the importance of using a similarity measure based on a likelihood ratio measure, as reported in [8, 3].[2]

2. A good estimation of the **normalization factor**, as discussed below.

As for the case of discriminant approaches in speech recognition [2], one of the difficulties in the development of speaker verification systems is thus the estimation of the normalization factor $P(X|\overline{S_c})$. Several solutions have been proposed. In one approach, we assume that the set of reference speakers already enrolled in the database is sufficiently representative of all possible speakers, and the normalization factor can then be estimated as

$$\log P(X|\overline{S_c}) \approx \sum_{S_i \in R, i \neq c} \log P(X|S_i) \tag{7}$$

where $R$ represents the set of speakers already enrolled in the system. One can also assume that the sum in (7) is dominated by the closest rival speaker, yielding the approximation

$$\log P(X|\overline{S_c}) \approx \max_{S_i \in R, i \neq c} \log P(X|S_i) \tag{8}$$

These solutions are, however, not often practical since:

1. In both cases, it will be necessary to estimate the conditional probabilities for all the reference speakers, which will often require too much computation.

2. In (8), the value of the maximum conditional probability varies a lot from speaker to speaker, depending on how close the nearest reference speaker is to the test speaker.

An alternative solution comprises considering a well chosen subset of reference speakers, usually called "cohort", on which $P(X|\overline{S_c})$ will be estimated. The cohort is usually defined as the group of speakers whose models are determined to be closed to or more "competitive" with the model of the target speaker $S_c$ [8, 16, 3]. A different cohort is thus assigned to every speaker and is automatically determined during the enrollment phase; it could also eventually be updated during the enrollment of new users. For this approach, the following approximation is used:

$$\log P(X|\overline{S_c}) \approx \sum_{S_i \in R_c} \log P(X|S_i) \tag{9}$$

where $R_c$ represents the cohort associated with speaker $S_c$. Experimental results show that this kind of normalization improves speaker separability and reduces the sensitivity to the decision threshold. In the spirit of a better approximation to (1), it was recently shown in [11] that it can be advantageous to include the model of the hypothesized speaker in the cohort. This improves the behavior of the algorithm in the cases where the acoustics for the actual speaker are rather different from the models for the claimed speaker identity (for instance, for the case of different gender), resulting in very small and unreliable likelihoods.

[1] The hypothesized speaker could also be included in the sum. This sometimes yields better estimates and better performance.

[2] In addition to normalizing scores, the likelihood ratio will also reduce the effect of some parameters affecting the similarity measure, such as the variability due to differences in transmission channel, as well as change in the speaker voice over time.

When HMMs are used, another solution consists of approximating $P(X|\overline{S_c}) \approx P(X|M)$, where $M$ is a speaker independent model that is trained either on a large set of speakers or only on the set of reference speakers. Depending on whether the verification system is text-dependent or text-independent, $M$ will either be a fully connected (ergodic) Markov model or a model representing the sentence to be pronounced.

Another solution could also consist of decomposing the problem of speaker/non-speaker discrimination into a series of 2-class problems. In [6], this problem was addressed by using many decision trees for each speaker $S_c$, each tree dealing with the problem of discriminating between $S_c$ and one specific speaker in the cohort (although the cohort was the same for every speaker). The set of decision trees can then be used to approximate $P(X|\overline{S_c})$. This approach was also shown to be somewhat less sensitive to the optimal setting of the decision threshold.

Following [5], we now briefly discuss some of the main speaker verification approaches.

## 4 Text-dependent speaker verification

In text dependent speaker verification, the system knows in advance the access password (or sentence) that will be used by the user. For each individual, there is a model that encodes both the speaker characteristics as well as the lexical content of the password. In this case, the techniques used for speaker verification are particularly similar to the methods used in speech recognition, namely:

1. Dynamic Time Warping (DTW) approach: in this case, the password of each user is simply represented as a small number of acoustic sequence templates corresponding to pronounciations of the password. During verification, the score associated with a new utterance of the password is computed via dynamic programming (and dynamic time warping) against the reference model(s). This approach is simple and requires relatively little computational resources during enrollment. It has been the basis of several commercial products.

2. HMM approach: in this case, the password associated with each user is represented by an HMM whose parameters are trained from several repetitions of the password. The amount of training required depends on the number of parameters, which can be a practical problem for larger models. Finally, the score associated with a new utterance is computed by either of the methods usually used in speech recognition: the Viterbi algorithm, which finds the best state path, or the forward ($\alpha$) recurrence, taking all possible paths into account. As generally found for speech recognition, HMM approaches have generally been found to be more accurate than simple DTW, but at the cost of higher computational requirements during training [17].

Given either of these approaches, a putative speaker identity can be verified using the similarity measure defined in Section 3 and comparing it to a decision threshold.

## 5 Text-independent speaker verification

In the case of text independent speaker verification, the lexical content of the utterance used for verification cannot be predicted. Since it is impossible to model all possible word sequences, different approaches have been proposed, including:

1. Methods based on long-term statistics such as the mean and variance calculated on a sufficiently long acoustic sequence. However, these statistics are a minimal representation of spectral characteristics, and can also be sensitive to the variability of the transfer function of the transmission channel.

   More recently, an alternative approach has been proposed in which the statistics of dynamic variables (e.g., in the cepstral domain) are used and modeled by a multi-dimensional autoregressive (AR) model [12]. In [7], different distance measures are compared for this AR approach, and it is shown that performance similar to standard HMM approaches can be achieved. It is also shown that the optimal order of the autoregressive process is around 2 or 3. Furthermore, correct normalization of the scores according to an *a posteriori* criterion seems essential to good performance.

2. Methods based on vector quantization. Vector quantization of spectral or cepstral vectors can be used to replace the original vector with an index to a codebook entry. In the case of speaker recognition, the spectral characteristics of each speaker can be modeled by one or more codebook entries that are representative of that speaker; see, for example, [19] for a typical reference. The score associated with an utterance is then defined as the sum of the distances between each acoustic vector in the sequence and its closest prototype vector from the codebook associated with the putative speaker (or codebooks associated with the cohort, for the normalization score). It is also possible to use a pitch detector and to define two sets of prototypes per speaker, one set each for voiced and unvoiced segments. For the voiced segments, pitch can then be added to the feature set to define the prototypes and compute the distance, requiring a choice of weights for the features.

   Finally, an alternative to "memoryless" vector quantization (so-called since each vector is quantized independently of its predecessors) was proposed in [9], in which source coding algorithms were used.

3. Fully connected (ergodic) HMMs. In this case, a fully connected HMM is trained during (according to Viterbi or Forward-Backward training) the enrollment of each user. The HMM states can then be defined in a completely arbitrary and unsupervised manner; in this approach, distances are stochastic and trained, but otherwise the approach is similar to the determination of codebook entries in vector quantization. Alternatively, states can be associated with specific classes; e.g., phones or even coarse phonetic categories. Some temporal constraints will generally be included in the models, typically by introducing minimum duration constraints on each state. Finally, several solutions using different topologies, different probability density functions associated with each state, as well as different training criteria, have been proposed, including:

   - HMMs trained according to a maximum likelihood criterion and having several (single or multi-) Gaussian states, or just a single multi-Gaussian state [11, 15]. Some discriminant training approaches typically used in speech recognition (e.g., Maximum Mutual Information [2]) have also been used in speaker verification to improve discrimination between users.
   - Autoregressive HMMs: In this case, the probability distribution associated with each state is estimated via an autoregressive process. Initially introduced by Poritz [14], this approach has been used with success by several laboratories [18]. Later on, this approach was also generalized to the class of HMMs using mixtures of autoregressive processes [20].

4. Artificial neural networks: Multilayer perceptrons have also been tested on speaker verification problems [13]. A specific neural network that has one or two output units is associated with each speaker. The weights of each network are trained positively using utterances from the corresponding speaker, and negatively on many utterances from rival speakers.

## 6 Text prompted speaker verification

Since verification is based on both the speaker characteristics and the lexical content of a secret password, text dependent speaker verification systems are generally more robust than text independent systems. However, both kinds of systems are susceptible to fraud, since for typical applications the voice of the speaker could be captured, recorded, and reproduced. In the case of a text-dependent system, even a password could be captured. To limit this risk, speaker verification systems based on prompted text have been developed. In this case, for each access, a recorded or synthetic prompt will ask the user to pronounce a different random sentence [3, 11]. The underlying lexicon could either be very large, or even be limited to the 10 digits, which would then be used to generate random digit strings. The advantage of such an approach is that impostors cannot predict the prompted sentence. Consequently, pre-recorded utterances from the customer will be of no use to the impostor. During each access, the system will prompt the user with a different sentence and a speech recognition system will be used prior to verification to validate the utterance. Finally, even when the utterance is rejected, the user can still be prompted with an additional sentence. Since the new sentence will be different, the acoustic vector sequence will not be too correlated with the previous one, which will improve the quality of the estimators by accumulating uncorrelated evidence. This strategy would not be as useful in a text dependent system, since the repeated sentence would have the same lexical content as the original one.

The speech recognition that is used before text-prompted verification is often based on phonetic HMMs, typically using Gaussian or multi-Gaussian distributions. These models are defined to cover the lexicon, and are independently trained on each user. A key difficulty with this approach is that there is typically not much enrollment data available to train the HMMs. For this reason, single-Gaussian single-state phonetic models are often used. Given the enrollment data generally available in speaker verification problems, such simple models have often performed as well as more complex models [3].

During verification, the system knows the prompted sentence and, using the phonetic transcription of the lexicon, can build the associated HMM model by simple concatenation of the constituent phones. The resulting model is then used to first validate the utterance (by computing the confidence level associated with the acoustic vector sequence) and then perform speaker verification. Given score normalization, a similar procedure can be used for the cohort speakers.

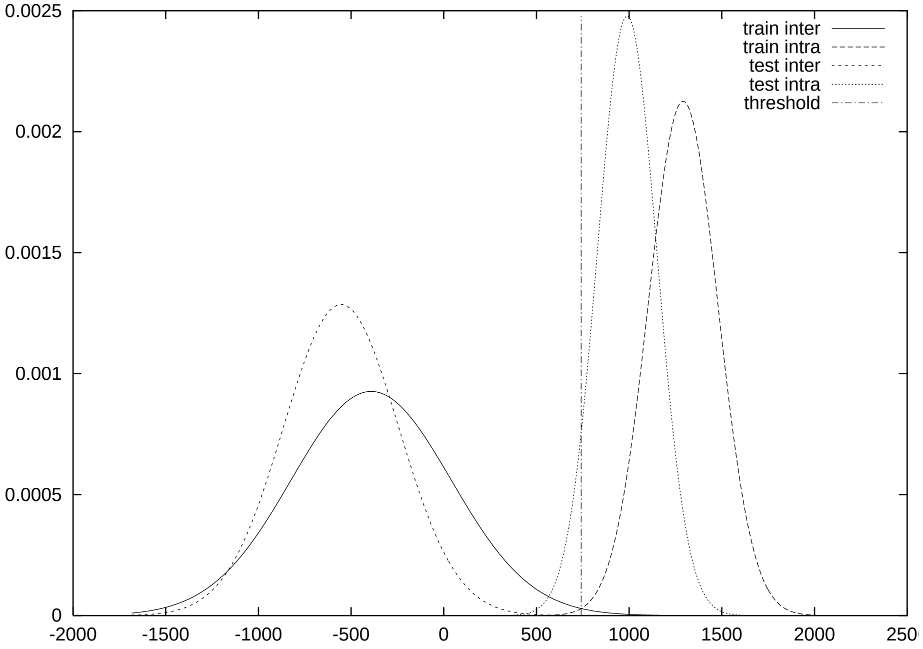

Figure 1: *Example of Gaussian approximations of the distributions of* $P(X|\overline{S_c})$ *(the left two Gaussians, respectively for training and test set) and* $P(X|S_c)$ *(right Gaussians). The vertical line (750) represents the decision threshold corresponding to the Equal Error Rate (EER) as estimated on the training data. As shown here, the position of the Gaussian can vary from training to test data, depending on the variability of channel and speaker characteristics. This illustrates the importance of using normalized scores, as discussed in Section 3. Means and variances were computed on a set of real data corresponding to a specific speaker* $S_c$ *and a given set of impostors.*

# 7 Identification, verification and the decision threshold

Each model discussed above can be used to compute a matching score (or a likelihood ratio) between some speaker model and a speech utterance. In the case of speaker identification, these scores are computed for all possible reference models and the identified speaker will be recognized as the one yielding the best score. In the case of speaker verification, scores are only computed for the putative speaker model and (when a normalization as discussed in Section 3 is used) the cohort models.

As illustrated by Figure 1, the resulting (normalized) score will be compared to a threshold above which the speaker will be accepted. Estimation of the optimal threshold is often critical in good performance of the system [8]. If the decision threshold is too high, too many customers will be rejected as impostors; such an error is referred to as a **false rejection**. Such a threshold will screen impostors very well, but at the cost of a high customer rejection rate. If the threshold is too low, too many impostors will be accepted as customers; this kind of error is called a **false acceptance** or (in more general signal detection parlance) a **false alarm**. Such a threshold will accept customers with little difficulty, but at the cost of a high impostor acceptance rate. In any real task, the cost of these two kinds of errors must be assessed for the real application in order to evaluate the utility of the system. As a convenience for system comparisons, the performance of speaker verification systems is often measured in terms of **equal error rate**, corresponding to the decision threshold where the false rejection rate is equal to the false acceptance rate.[3] This EER threshold is often estimated by assuming that the two distributions of $P(X|S_c)$ and $P(X|\overline{S_c})$ (i.e., fitting two Gaussians on experimental points, obtained for both the customer and a set of rival speakers) are Gaussian and computing the resulting EER point. If this EER threshold is computed on the training set, it will be referred to as **a priori threshold**, while it will be referred to as **a posteriori threshold** is computed on the test set. Of course, in real systems, the a posteriori EER measure will not be accessible (since in any one application the system operates with some particular scheme for setting the threshold), but EER is often approximated as half of the sum of the two error rates.

## References

[1] Atal, B.S., Automatic speaker recognition based on pitch contours, *Journal of Acoustical Society of America*, vol. 52, pp. 1687-1697, 1972.

[2] Bahl, L.R., Brown, P.F., de Souza, P.V., and Mercer, R.L., Maximum mutual information estimation of hidden Markov model parameters, *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Tokyo), pp. 49-52, 1986.

[3] de Veth, J. & Bourlard, H., Comparison of Hidden Markov Model Techniques for Automatic Speaker Verification in Real-World Conditions, *Speech Communication* (North-Holland), vol. 17, no. 1-2, pp. 81-90, 1995.

[4] Doddington, G., Speaker recognition-identifying people by their voices, *Proceedings of the IEEE*, vol. 73, pp. 1651-1664, 1985.

[5] Furui, S., An overview of speaker recognition technology, in *Automatic Speech and Speaker Recognition*, C.-H. Lee, F.K. Soong, K.K. Paliwal (eds.), pp. 31-56, Kluwer Academic Publishers, 1996.

[6] Genoud, D., Moreira, M., and Mayoraz, E., Text dependent speaker verification using binary classifiers, *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Seattle, Washington), pp. 129-132, 1998.

[7] Griffin, C., Matsui, T., & Furui, S., Distance measures for text-independent speaker recognition based on MAR model, *Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Processing* (Adelaide, Australie), pp. I-309-312, 1994.

[8] Higgins, A.L., Bahler, L., & Porter, J., Speaker verification using randomized phrase prompting, *Digital Signal Processing*, vol. 1, pp. 89-106, 1991.

[9] Juang, B.-H. & Soong, F.K., Speaker recognition based on source coding approaches, *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Albuquerque, NM), pp. 613-616, 1990.

[3] 1998 speaker verification systems are reporting EER performance varying between 0.1% and 5%, depending on the conditions.

[10] Matsui, T. & Furui, S., Text-independent speaker recognition using vocal tract and pitch information, *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Albuquerque, NM), pp. 137-140, 1990.

[11] Matsui, T. & Furui, S., Concatenated phoneme models for text-variable speaker recognition, *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Minneapolis, MI), pp. II-391-394, 1993.

[12] Montacie, C., Deleglise, P., Bimbot, F., & Caraty, M.-J., Cinematic techniques for speech processing: temporal decomposition and multivariate linear prediction, *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (San Francisco, CA), pp. I-153-156, 1992.

[13] Oglesby, J. & Mason, J.S., Optimization of neural models for speaker identification, *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Albuquerque, NM), pp. 261-264, 1990.

[14] Poritz, A.B., Linear predictive hidden Markov models and the speech signal, *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Paris, France), pp. 1291-1294, 1982.

[15] Rose, R.C. & Reynolds, R.A., Text independent speaker identification using automatic acoustic segmentation, *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing,* (Albuquerque, NM), pp. 293-296, 1990.

[16] Rosenberg, A.E., DeLong, J., Lee, C.-H., Juang, B.H., and Soong, F.K., The use of cohort normalized scores for speaker verification, *Proc. Intl. Conf. on Spoken Language Processing*, pp. 599-602, 1992.

[17] Rosenberg, A.E., Lee, C.-H., & Gokcen, S., Connected word talker verification using whole word hidden Markov models, *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Toronto, Canada), pp. 381-384, 1991.

[18] Savic, M. & Gupta, S.K., Variable parameter speaker verification system based on hidden Markov modeling, *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Albuquerque, NM), pp. 281-284, 1990.

[19] Soong, F.K., Rosenberg, A.E., Rabiner, L.R., & Juang, B.-H., A vector quantization approach to speaker recognition, *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Tampa, FL), pp. 387-390, 1985.

[20] Tishby, N.Z., On the application of mixture AR hidden Markov models to text independent speaker recognition, *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-30, pp. 563-570, 1991.