Comparative study of speaker verification techniques based on vector quantization, sphericity models and dynamic time warping*

Sofia Tsekeridou† C. Kotropoulos† A. Xafopoulos† I. Pitas†

Abstract — **Three simple speaker verification techniques based on vector quantization, sphericity models and dynamic time warping, respectively, are developed and tested using the same experimental protocol. Two types of feature vectors, the linear prediction derived cepstral coefficients and the mel-frequency cepstral coefficients are considered. The efficiency of the combination of the type of acoustic analysis and the verification technique is quantitatively measured through the achieved equal error rate.**

1 Introduction

Three simple speaker verification techniques are developed and tested using the same experimental protocol. The first two methods are text-independent ones. They are based on vector quantization (VQ) [2, 3] and sphericity-based models [4]. Text-independent verification is often preferred over a text-dependent one, because it maximizes user comfort [5]. The third method is text-dependent and employs dynamic time warping (DTW) [6]. The chosen methods do not require large training databases in order to build the reference models needed to parameterize the speech or the speakers. This is not the case with other speaker verification techniques, such as Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs). Moreover, they have limited storage requirements. Indeed, VQ and sphericity-based models possess much less storage requirements than DTW, HMMs and GMMs. The storage requirements of HMMs depend on the number of states used to model each word, the number of components in the Gaussian mixture describing each state, and the type of the covariance matrix (i.e., full or diagonal) employed in the multivariate Gaussian probability density function of each component. It can be shown that DTW has less storage requirements than HMMs when many states per word and full covariance matrices are employed in the latter. In this work a fixed-vocabulary speaker verification application (i.e., a vocabulary composed of digits 0-9 in French) is explored. Two types of acoustic signal analysis are investigated, namely the linear prediction (LP) derived cepstral coefficients (LPCCs) and the mel-frequency cepstral coefficients (MFCCs).

The major contribution of the paper is in the comparative study of the efficiency of the combination of the acoustic signal analysis performed and the speaker verification algorithm employed. The efficiency of the combination of the type of acoustic analysis performed and the verification technique employed is quantitatively measured through the achieved equal error rate (EER) using the same experimental protocol on the M2VTS database. This database has been selected instead of other speech databases (e.g., TIMIT, Polycost or OGI) so that the verification rates reported here can be subsequently fused with those obtained using frontal face images toward a multi-modal access control system, as in [7]. The same database has also been used in [7, 8]. There, experimental results have primarily been reported for audio-based speaker authentication using HMMs. However, the reported rates refer only to LPCCs and to one of the experiments described in this paper. The same remarks apply to the rates reported for sphericity-based speaker verification in [9].

A second contribution is the use of Learning Vector Quantizer[1] as a speaker verification method based on VQ. The advantage of the method is that it allows the parallel generation of speaker-dependent codebooks. Furthermore being a supervised clustering algorithm, it penalizes wrong classifications of feature vectors to speaker-dependent classes maximizing thus the inter-speaker separation. Intra-speaker variability is taken into account by employing Mahalanobis distance measures. Another extension of the basic algorithm is the use of segment-based distances in both the training and the recall phase. That is, the distances are measured on groups of successive feature vectors that belong to the same speech unit (e.g., digit in our case).

*This work was supported by the European Union Research Training Network "Multi-modal Human-Computer Interaction" (HPRN-CT-2000-00111).

†Dept. of Informatics, Aristotle Univ. of Thessaloniki, Box 451, Thessaloniki 54006, GREECE. E-mail: {sofia,costas,alexandr,pitas}@zeus.csd.auth.gr, Tel. +30-31-998225, Fax: +30-31-998419.

[1] LVQ3 according to the terminology used in [10].

2 Audio Preprocessing and Feature Vector Extraction

To discard the silent parts from the utterance an endpoint detection algorithm that uses short-term measures of energy and zero-crossing rate is employed [11]. Subsequently, the speech frames are pre-emphasized. Preemphasis is performed by filtering a speech frame with an FIR filter having transfer function $H(z) = 1 - 0.95z^{-1}$ in order to increase the relative energy of its high-frequency spectrum.

Two types of feature vectors parameterize the speech frames that have been uttered by each speaker: (a) LPCCs, and (b) MFCCs. In order to evaluate the LPCCs of a speech frame, linear prediction analysis of order p is performed using the autocorrelation method [3, 11]. The LPCCs have undergone liftering to increase their robustness, as is proposed in [1]. We have used frames of duration 30 ms with an overlap of 20 ms between successive frames and a model order $p = 12$. In contrast to [9] the log signal energy is not included in the feature vector. Moreover, delta and delta-delta coefficients are not considered. MFCCs are of the most popular feature vectors derived from the acoustic analysis [12]. Their computation was done according to the procedure described in [3, 11] for a decomposition of the Nyquist bandwidth into $L = 40$ triangular filters.

3 Verification based on variants of Learning Vector Quantizer

Speaker-dependent codebooks are generated in parallel for each authorized user (i.e., *client*) of the system. Each speaker-dependent codebook is comprised of an arbitrarily chosen number of codevectors (e.g., 32). The number of codevectors is usually chosen so that it approximates the number of phones in the language. Although French has 41 phones, to a first degree of approximation, we model the French consonants with 20 codevectors (i.e., 12 for obstruents and 8 for sonorants) and the French vowels with 12 codevectors by omitting the distinction between oral and nasal vowels [13, pp. 705-706].

Let N_{sp} be the total number of training speakers. Speaker dependent-codevectors are initialized by the LBG algorithm applied to the feature vectors. Each speaker is modeled by its covariance matrix $\mathbf{\Gamma}_i$ and its codebook $\mathcal{M}_i$, $i = 1, 2, \ldots, N_{sp}$. $\mathbf{\Gamma}_i$ is estimated from the feature vectors of the speaker's training utterance. During the training procedure, a feature vector $\mathbf{c}$ is quantized by the kth codevector $\mathbf{m}_{ik}$, $k = 1, 2, \ldots, 32$, of the speaker codebook $\mathcal{M}_i$, if

$$(i, k) = \arg \min_{\substack{j=1,2,\ldots,N_{sp} \\ l=1,2,\ldots,32}} d(\mathbf{m}_{jl}, \mathbf{c}) \tag{1}$$

where

$$d(\mathbf{m}_{jl}, \mathbf{c}) = (\mathbf{m}_{jl} - \mathbf{c})^T \mathbf{\Gamma}_j^{-1} (\mathbf{m}_{jl} - \mathbf{c}). \tag{2}$$

The update of codevectors $\mathbf{m}_{ik}$ is done as in LVQ3 [10].

Having modeled each speaker by a codebook $\mathcal{M}_i$, $i = 1, 2, \ldots, N_{sp}$ and a covariance matrix $\mathbf{\Gamma}_i$, during pattern matching, a test feature vector $\mathbf{c}_t$ is quantized by the codevector $\mathbf{m}_{ik}$ that yields the minimal Mahalanobis distance (2). A score function $\delta(\mathbf{m}_i, \mathbf{c}_t) = d(\mathbf{m}_{ik}, \mathbf{c}_t)$ is then computed. The average distortion from the entire speaker utterance

$$D_i = \frac{1}{T} \sum_{t=1}^{T} \delta(\mathbf{m}_i, \mathbf{c}_t) \tag{3}$$

where T is the total number of feature vectors extracted from the speaker's utterance is used as a global matching score.

Another extension of the LVQ3 algorithm has been considered. Usually, speech segments are chosen so that they correspond to speech units (i.e., digits in our case). Let us suppose that the total number of speech segments in the speaker's utterance is N_{seg}. Instead of updating the speaker codebooks for each training feature vector separately, the updating can be performed for the group of feature vectors $\mathbf{c}_l$, $l = 1, 2, \ldots, N_{c_r}$, that belong to the rth segment, $r = 1, 2, \ldots, N_{seg}$, based on the minimization of segment-based distances

$$\bar{\delta}(\mathbf{m}_i, \{\mathbf{c}_l, l = 1, 2 \ldots, N_{c_r}\}) = \frac{1}{N_{c_r}} \sum_{l=1}^{N_{c_r}} \delta(\mathbf{m}_i, \mathbf{c}_l) \tag{4}$$

where N_{c_r} is the total number of consecutive feature vectors in the rth segment. During pattern matching, the average distortion D_i is estimated by

$$D_i = \frac{1}{N_{seg}} \sum_{r=1}^{N_{seg}} \bar{\delta}(\mathbf{m}_i, \{\mathbf{c}_l, l = 1, \ldots, N_{c_r}\}). \tag{5}$$

4 Verification using sphericity-based models

Each client is modeled by the covariance matrix $\mathbf{\Gamma}_X$ of the feature vectors of the client's training utterances. Similarly, a test person is modeled by the covariance matrix $\mathbf{\Gamma}_Y$ of the feature vectors

of its utterance. The similarity measure between the client and the test person is the arithmetic-harmonic sphericity measure $D_{SPH}(\mathbf{\Gamma}_X, \mathbf{\Gamma}_Y)$ given by [4]:

$$D_{SPH}(\mathbf{\Gamma}_X, \mathbf{\Gamma}_Y) = \log \frac{\mathrm{tr}(\mathbf{\Gamma}_Y^{-1}\mathbf{\Gamma}_X)\mathrm{tr}(\mathbf{\Gamma}_X^{-1}\mathbf{\Gamma}_Y)}{q^2} \quad (6)$$

where q is the feature vector dimensionality. In our analysis, $q = p = 12$.

5 Verification based on Dynamic Time Warping

In DTW, a speaker-dependent template model is created for each digit. The template model is comprised of the parameterized speech frames that correspond to the digit under study $(\mathbf{c}_1(X), \mathbf{c}_2(X), \ldots, \mathbf{c}_N(X))$. The parameterized speech frames of a test person that utters the same digit form a sequence $(\mathbf{c}_1(Y), \mathbf{c}_2(Y), \ldots, \mathbf{c}_M(Y))$. The latter sequence is aligned through a dynamic programming procedure by which temporal regions of the test person utterance are matched with appropriate regions of the template model [11]. A match score is derived per digit, i.e.,

$$D_{DTW}(X, Y) = \sum_{i=1}^{M} d(\mathbf{c}_i(Y), \mathbf{c}_{j(i)}(X)) \quad (7)$$

where $j(i)$ is the index of the reference feature vector that matches best the test feature vector indexed by i. All match scores are then summed to yield a global distance measure. In (7), $d(\mathbf{c}_i(Y), \mathbf{c}_{j(i)}(X))$ can be a Euclidean distance, since the feature vectors consist of cepstral coefficients.

6 Performance evaluation and discussion

The proposed speaker verification algorithm has been tested on the M2VTS database [14]. Five recordings of the 37 persons have been collected. Let BP, BS, CC, ..., XM be the identity codes of the persons in the database. Four experimental sessions have been implemented using a combination of the "leave-one-out" principle and rotation estimates with the first four recordings.

First let us describe the training procedure when the utterance of person BP from the fourth recording is left out so that BP is used as test impostor, and similarly the utterances of the remaining 36 persons $(BS, CC, \ldots, XM)$ from the same recording are also excluded so that these utterances are used to form test client claims. Accordingly the training set is built of the three recordings of utterances of the 36 clients, $BS, CC, \ldots, XM$. Six permutations of the three recordings taken two at a time can be made. In each permutation, the parameterized speech frames of utterances from the first recording are used to build the speaker-dependent models for all clients. Using the client models and the utterances from the second recording one may compute: (i) 1 distance measure between the model and utterance of each client, and, (ii) 35 distance measures between the model of each client and the utterance of any different speaker. To illustrate the derivation of thresholds, let us consider what happens when person BP pretends to be person BS using his utterance from the fourth recording. Let $D_{(l)}(BS, (BP, 4))$ denote the lth order statistic in the set of 35 distance measures between a model of BS and utterances of $CC, \ldots, XM$ that constitute the training impostor claims for BS using the training set, as is described previously. A threshold for person BS can by chosen as:

$$T_{BS}(BP, 4) = D_{(l)}(BS, (BP, 4)), \quad l = 1, 2, \ldots \quad (8)$$

A test procedure is then defined where the first three recordings create the training set while the fourth one is used as a test set. Each person of the test set is considered in turn as an impostor while the 36 others are used as clients. Each client tries to access under its own identity while the impostor tries to access under the identity of each of the 36 clients in turn. The speaker-dependent models and thresholds derived during the training procedure are exploited. Since the training procedure delivers three speaker-dependent models, we compute the minimum distance measure between the test utterance and the speaker-dependent models. The minimum distance is then compared against the threshold given by (8). By rotating between the four recordings, 5328 client and additional 5328 impostor claims can be produced. For a particular choice of parameter l, a collection of thresholds is determined that defines an operating state of the test procedure. For such an operating state, a *false acceptance rate* (FAR) and a *false rejection rate* (FRR) can be computed. We may create a plot of the FRR versus the FAR with the scalar ϱ as a varying parameter, the so-called Receiver Operating Characteristic (ROC) of the verification technique. Although the ROC curves have been derived due to lack of space they are not included in the paper. To assess better the performance of the verification algorithms, we use the fifth recording as a testbed and the first three recordings to train each verification algorithm, as in [8]. The Equal Error Rates (EER) achieved in all cases are tabulated in Table 1. Note that DTW yields a first operating point at FAR=1.53% and FRR=0.65% when

Table 1: Achieved EER values (%).

Classifier	LPCC		MFCC	
	Entire	Shot 5	Entire	Shot 5
LVQ3	7.98	6.35	4.52	4.39
LVQ3 on segments	7.23	6.97	4.36	3.68
Sphericity models	2.70	2.48	1.60	1.47
DTW	[0.65, 1.53]	2.7	4.18	5.4

LPCCs are employed. Therefore, an ERR cannot be estimated in this case. It is seen that MFCC feature vectors yield better results than LPCC feature vectors for text-independent speaker verification. On the contrary, LPCC feature vectors outperform MFCC feature vectors in the case of DTW (i.e., text-dependent speaker verification). This is in par to the observation that LPCC coefficients outperform MFCC coefficients for simple Hidden Markov Model (HMM) topologies [1]. For the text-independent speaker verification algorithms tested, a drop in EER is found when the fifth recording is processed. In the case of DTW, a larger EER is measured when the fifth recording is used for testing. This fact is attributed to the not optimal segmentation of the speech signal into digits. The use of training digit models, training silence models and HMMs to segment the speech signal into digits would solve the problem. A slightly better performance is obtained when segment-based distances are used in the variant of LVQ3 classifier both for LPCC and MFCC feature vectors. For comparison purposes we refer that HMMs using a parameterization of speech frames with LPCCs of order 12 including the log-energy coefficient, delta and delta-delta coefficient have given FAR=2.3% and FRR=2.8 % [8] when the fifth recording is used for test purposes. It can be seen that a similar performance level has been obtained by the sphericity-based models and DTW with LPCCs and a better performance has been obtained with sphericity-based models with MFCCs. The performance of LVQ3 on segments and MFCC parameterization is not far behind.

References

[1] F. Bimbot, M. Blomberg, L. Boves, D. Genoud, H.-P. Huttel, C. Jaboulet, J. Koolwaaij, J. Lindberg, and J.-B. Pierrot, "An overview of the CAVE project research activities in speaker verification," *Speech Communication*, vol. 31, pp. 155–180, 2000.

[2] F.K. Soong, A.E. Rosenberg, and B.-H. Juang, "A vector quantization approach to speaker recognition," *AT & T Technical Journal*, , no. 66, pp. 14–26, 1987.

[3] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[4] D. Genoud, G. Gravier, F. Bimbot, and G. Chollet, "Combining methods to improve the phone-based speaker verification decision," in *Proc. ICSLP'96*, 1996, vol. 3, pp. 1756–1759.

[5] J. de Veth and H. Bourlard, "Comparison of hidden markov model techniques for automatic speaker verification in real-world conditions," *Speech Communication*, vol. 17, pp. 81–90, 1995.

[6] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[7] B. Duc, G. Maître, S. Fischer, and J. Bigun, "Person authentication by fusing face and speech information," in *Audio- and Video-based Biometric Person Authentication*, J. Bigün, G. Chollet, and G. Borgefors, Eds., vol. 1206 of *Lecture Notes in Computer Science*, pp. 311–318. Springer, Berlin, 1997.

[8] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," in *Audio- and Video-based Biometric Person Authentication*, J. Bigün, G. Chollet, and G. Borgefors, Eds., vol. 1206 of *Lecture Notes in Computer Science*, pp. 318–326. Springer, Berlin, 1997.

[9] G. Maître, S. Ben-Yacoub, and J. Luettin, "Two voice-based authentication algorithms," in *Multimodal Verification Techniques and Evaluation*, J. Bigün and Y. Abdeljaoued, Eds. 1998.

[10] T. Kohonen, "The self-organization map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[11] J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*, Prentice Hall, Upper Saddle River, NJ, 1993.

[12] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[13] D. Gibbon, R. Moore, and R. Winski, Eds., *Handbook of Standards and Resources for Spoken Language Systems*, vol. IV, Mouton de Gruyter, Berlin, 1998.

[14] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database," in *Audio-and Video-based Biometric Person Authentication*, J. Bigün, G. Chollet, and G. Borgefors, Eds., pp. 403–409. Springer, Berlin, 1997.