AR-VECTOR USING CMS FOR ROBUST TEXT INDEPENDENT SPEAKER VERIFICATION

Charles B. de Lima[1]*, *Dirceu G. da Silva*[1], *Abraham Alcaim*[2], *and José A. Apolinário Jr.*[1]

[1] IME - Department of Electrical Engineering, Praça General Tibúrcio, 80—Urca, 22.290-270 Rio de Janeiro, RJ, Brazil
cborges/dirceu/apolin@epq.ime.eb.br
[2]CETUC/PUC-Rio, Rua Marquês de São Vicente, 225—Gávea, 22453-900, Rio de Janeiro, RJ, Brazil
alcaim@cetuc.puc-rio.br

Abstract: This paper presents the performance of the AR-Vector with Cepstral Mean Subtraction (CMS) used to compensate the distortions caused by distinct telephone channels. The performance obtained with the use of CMS is compared with a system without compensation. With $60s$ of speech signal used for training and $30s$ used for testing, the error rate without channel normalization is around 2.82% against the 1.65% achieved with CMS. For $10s$ testing time, the error rate dropped from 5.40% to 3.80% when using CMS. For the lowest testing time ($3s$), the error rate of the AR-Vector is close to 19% regardless the use or not of the normalization technique. Although there is a clear improvement in performance when using CMS, it is not of major significance. This leads to the conclusion that the AR-Vector classification system is somewhat robust to channel distortion, especially as the testing time decreases.

1. INTRODUCTION

The recognition of a human being through his voice is one of the simplest forms of automatic recognition because it uses biometric characteristics which come from a natural action, the speech. Speech, being present everywhere from telephone nets to personal computers, may be the cheapest form to supply a growing need of providing authenticity and privacy in the worldwide communication nets [1].

Research in the area of speaker recognition has significantly grown over the last few years due to a vast area of applications where the recognition can be used. Some of these applications are as follows.

– Access control: devices, networks, and data in general;

– Authentication for business transactions as a tool to prevent fraud in: shopping over telephone, credit card validation, transactions over Internet, bank operations, etc.

– Law enforcement: penitentiary monitoring, forensic applications, etc.

– Help to handicapped.

– Military use: classified information requiring speaker verification.

The speech for security purpose can be used with other validation devices such as magnetic cards and passwords. It is expected that in the future more and more applications include man machine iteration: e-mail being used by everyone and speech operated devices controlling the sound and the illumination of public environments and cars.

Speaker verification is the task of verifying if a speech signal (utterance) belongs or not to a certain person, which means a binary decision. The decisions are carried out in the so-called speakers open set [2] because the recognition is done in an unknown speakers set (possible impostors). As to text dependency, recognition can be dependent or independent. Systems demanding a pre-determined word or phrase are text dependent. Such systems can offer precise and reliable comparisons between two speech signals with the same text, in phonetically similar environments, requiring only 2 to 3 seconds of speech for training and testing. In text independent systems, such comparisons are not so easy to be obtained. The performance decreases as compared to text dependency. Moreover, in order to obtain reasonable statistics of the signal, it is, in general, necessary from 10 to 30 seconds of speech signal for training and testing [3].

The AR-Vector is a model able to capture informations about the dynamics of a given speaker, interpreted as the speaker articulatory capacity or, in other words, the way he or she speaks as time goes by [4].

* The author thanks CAPES, Brazil, for partial funding of this work.

The AR-Vector can also be seen as an extension of a very well known model used in speech processing, the Linear Predictive Coding (LPC). Whilst the LPC is based on the linear regression over scalars, AR-Vector is based on the regression over feature vectors. For an evaluation, AR-Vector needs a distance measure in order to compare two models. For this distance, it is usually employed the so-called Itakura distance [5].

The use of Cepstral Mean Subtraction minimizes the effect of a transmitting channel; such technique is widely used for channel normalization [6]. In this work we evaluate its use in AR-Vector models, when the training and testing telephone channels exhibit distinct characteristics.

This paper is organized as follows. In Section 2, the AR-Vector is reviewed. Section 3 contains the details of the system configuration used in our experimental procedure followed by simulation results in Section 4. Finally, concluding remarks are presented in Section 5.

2. THE AR-VECTOR MODEL

The AR-Vector is actually an extension of the LPC in the sense that it carries out a prediction among vectors (not samples), modeling the time evolving of the vectors (in our case, the feature vectors of speech). The order p AR-Vector model for a sequence of N vectors of dimension $m \times 1$, in time domain, is given by:

$$\mathbf{X}_n = \sum_{k=1}^{p} \mathbf{A}_k X_{n-k} + \mathbf{E}_n \tag{1}$$

where $\mathbf{X}_n$ and $\mathbf{E}_n$ are dimension $m \times 1$ vectors, with $\mathbf{E}$ representing the linear prediction error, and $\mathbf{A}_k$ being an $m \times m$ prediction matrix. The set of prediction matrices can be represented by an $m \times (p+1)$ matrix $\mathbf{A} = [\mathbf{A}_0 \quad \mathbf{A}_1 \quad \mathbf{A}_2 \quad \cdots \quad \mathbf{A}_p]$, with $\mathbf{A}_0$ being the identity matrix or $\mathbf{A}_0 = \mathbf{I}$.

From the vectors $\mathbf{X}_n$, we can define an estimate of the autocorrelation matrix:

$$\mathbf{R}_k = \sum_{n=0}^{N-k} \mathbf{X}_n \mathbf{X}_{n+k}^T \tag{2}$$

where N is the number of vectors $\mathbf{X}_n$ available for the estimation. Note that $\mathbf{R}_k$ results in a $m \times m$ matrix.

$\mathbf{A}_k$ are obtained by solving the following set of equations.

$$\begin{pmatrix} \mathbf{R}_0 & \mathbf{R}_1^T & \dots & \mathbf{R}_{p-1}^T \\ \mathbf{R}_1 & \mathbf{R}_0 & \dots & \mathbf{R}_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{p-1} & \mathbf{R}_{p-2} & \dots & \mathbf{R}_0 \end{pmatrix} \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_p \end{pmatrix} = \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \\ \vdots \\ \mathbf{R}_p \end{pmatrix} \tag{3}$$

From the previous equation, if we define the $pM \times pM$ Toeplitz autocorrelation matrix as $\boldsymbol{\mathcal{R}}$, the $pM \times M$ coefficient matrix as $\boldsymbol{\mathcal{A}}$, and the $pM \times M$ autocorrelation matrix on the right-hand side as $\mathbf{R}$, we have:

$$\boldsymbol{\mathcal{R}}\boldsymbol{\mathcal{A}} = \mathbf{R} \quad \Rightarrow \quad \boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{R}}^{-1}\mathbf{R} \tag{4}$$

Once $\mathcal{R}$ is a Toeplitz matrix, a well known computationally efficient algorithm (the Levinson-Durbin recursion) can be used to solve the set of equations [7].

The utilization of the AR-Vector in speaker recognition requires the use of some measure to evaluate the similarity between two autoregressive models. A widely used distance measure is the Itakura distance [5] which provides the distance between two all-poles LPC's based on the linear prediction coefficients and on the autocorrelation matrix.

The use of the Itakura distance with the AR-Vector is presented in [4]. Assuming a stored model $\boldsymbol{\mathcal{A}}$ previously estimated from a given speaker and a model $\boldsymbol{\mathcal{B}}$ from a pretense speaker, three distance measures between these two model are defined for their respective autocorrelation matrices. These measures are:

1. Distance from $\boldsymbol{\mathcal{B}}$ to $\boldsymbol{\mathcal{A}}$:

$$d(\boldsymbol{\mathcal{B}},\boldsymbol{\mathcal{A}}) = \log\left[\operatorname{tr}\left(\frac{\boldsymbol{\mathcal{A}}\boldsymbol{\mathcal{R}}_B\boldsymbol{\mathcal{A}}^T}{\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{R}}_B\boldsymbol{\mathcal{B}}^T}\right)\right] \tag{5}$$

2. Distance from $\boldsymbol{\mathcal{A}}$ to $\boldsymbol{\mathcal{B}}$:

$$d(\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{B}}) = \log\left[\operatorname{tr}\left(\frac{\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{R}}_A\boldsymbol{\mathcal{B}}^T}{\boldsymbol{\mathcal{A}}\boldsymbol{\mathcal{R}}_A\boldsymbol{\mathcal{A}}^T}\right)\right] \tag{6}$$

3. Symmetric Distance:

$$d_{\text{sym}} = \frac{1}{2}\left[d\left(\boldsymbol{\mathcal{B}},\boldsymbol{\mathcal{A}}\right) + d\left(\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{B}}\right)\right] \tag{7}$$

The speaker verification system provides a binary output, acceptance or rejection of a pretense speaker. Hence, an estimation of a threshold θ, based on true and false utterances, is required. This threshold is estimated with the *true distances*, i.e., the two models under comparison are from the same person, and with the *false distances* given by the pretense speaker model compared to the other models not belonging to him.

From these distances, the threshold is estimated taking into account false acceptance errors and false rejection errors. When a speaker is to be analyzed, he will be accepted if the resulting distance is lower than the threshold. He will be rejected otherwise. Fig. 1 presents the AR-Vector verification system.

The autoregressive model produces a smoothed model of the evolving features, capturing information from the dynamics of the speaker.

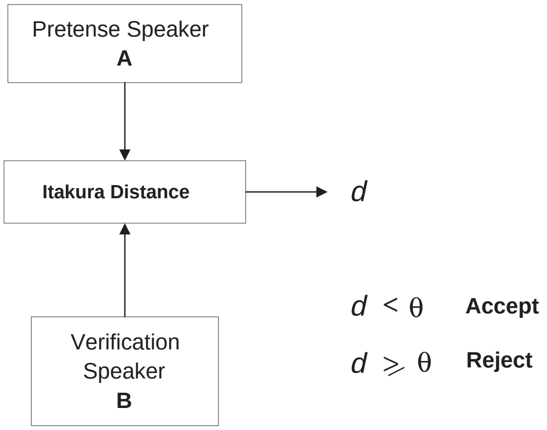

Fig. 1. AR-Vector Speaker Verification System.

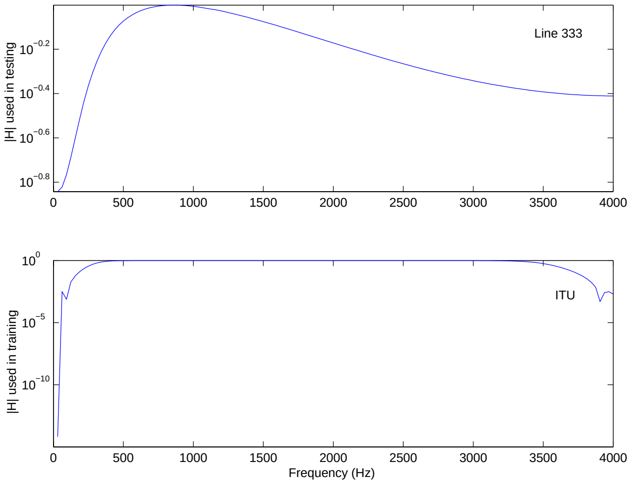

Fig. 2. Frequency responses of the telephone channels used for testing (upper curve) and training (lower curve).

3. EXPERIMENTAL PROCEDURE

The utterances used in our experiments were recorded with $8KHz$ as sampling rate, electret microphones, and in a low noise environment. We have used 40 male speakers. Each speaker uttered 200 sentences, in Brazilian Portuguese, extracted from [8]. We have used 15 mel-cepstrum coefficients (MCC) [9], with $20ms$ windows and 50% overlapping. The silence between words was eliminated.

In our experiments, the AR-Vector used order 2 with the symmetric Itakura distance (previous experiments have shown its better performance for this configuration). We have used 60, 30, and $10s$ of speech signal for training and 30, 10, and $3s$ for testing. The setting of the decision threshold was established in order to equally minimize the error rate between false acceptance—FA (to accept someone which does not correspond to the true speaker)—and false rejection—FR (to reject someone which corresponds to the true speaker). This procedure resulted in an equal error rate (EER) measure [2]. The training data were corrupted by a different channel than the testing data—see in Fig. 2 the frequency responses of the two telephone channels used in our experiments. The system was evaluated without channel normalization (WN) and with CMS normalization (CMS).

4. SIMULATION RESULTS

The EER results obtained with the order 2 AR-Vector using the symmetrical Itakura distance and 60s training are shown in Table. 1. From this table, we can see a 2.82% EER when using $30s$ for testing time without channel normalization. With CMS, the system performance improves to an EER of 1.65%. When the testing time decreases to $10s$, the error rate drops from 5.40% to 3.80% with the CMS normalization. On the other hand, when $3s$ of speech signal is used for testing, the AR-Vector model is not accurate due to the lack of data for the cepstral coefficient vectors modeling. For this reason, normalizing the channel presents no influence in the verification error rates. The resulting EER is close to 19% with or without CMS.

Table 1. Performance of the AR-Vector for $60s$ training.

System	tests(EER %)		
	30s	10s	3s
WN	2.82	5.40	18.84
CMS	1.65	3.80	18.90

The DET (Detection Error Tradeoff) [10] curves, shown in Fig. 3, yield the performance of the several system configurations for 60s training and 30, 10 and 3s testing. On these curves it is possible to choose the system operating point in terms of *false acceptance*(FA) and *false rejection*(FR) error rates according to the desired application.

We can clearly note that the amount of time used for training and testing has a strong influence on the results: the system performance improves when more data is available. The use of the normalization scheme (CMS) has proved to be effective when the signals (training and testing) are corrupted by different channels. However, the gains provided by the CMS rapidly decreases with the testing time. For 3s testing it yields no improvement for any operating point, as can be seen from Fig. 3.

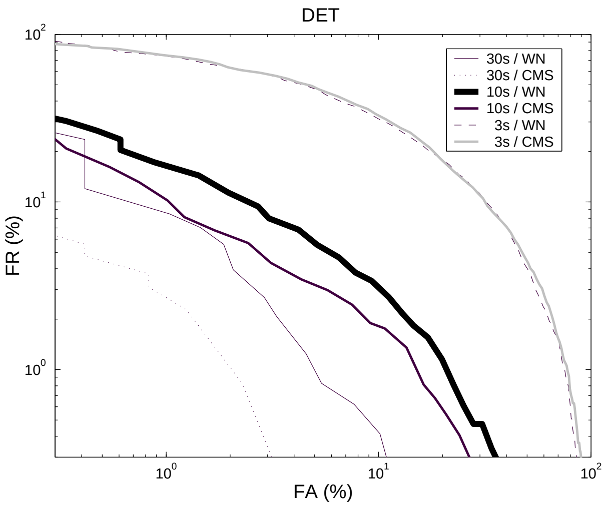

Fig. 3. The DET curves showing the performance of the AR-Vector for $60s$ of training and 30, 10 and 3s testing.

5. CONCLUSION

In this paper we have examined the performance of the AR-Vector model when using Cepstral Mean Subtraction for the task of speaker verification when the speech signals used for training and testing suffered from distortions dues to different channels. We have considered the ITU (Recommendation G.151) and the Line 333 (model of a poor continental channel) fixed telephone channels for training and testing, respectively. We have found that the use of CMS with AR-Vector provided a performance improvement that decreases as the testing time is reduced. We have also noticed that the AR-Vector seems to be a promising speaker verification scheme because it is somewhat robust in case of different channel distortions: the normalization used (CMS) was indeed effective but the results without normalization were not so far apart. This is probably due to the long term analysis typically carried out by the AR-Vector. Further studies concerning the effectiveness of this technique when noise is added to the channel distorted speech is currently under investigation.

6. REFERENCES

[1] CAMPBELL, Joseph P., Jr. *Speaker Recognition: A Tutorial.* Proceedings of the IEEE, vol. 85, number 9, pp. 1437-1462, September 1997.

[2] REYNOLDS, Douglas A. *Speaker Identification and Verification Using Gaussian Mixture Speaker Models.* Speech Communication, vol. 17, pp. 91-108, 1995.

[3] JAYANT M. Naik. *Speaker Verification: A Tutorial.* IEEE Communication Magazine, pp. 42-47, January 1990.

[4] BIMBOT, F., L. Mathan, A. de Lima, and G. Chollet. *Standard and Target Driven AR-vector Models for Speech Analysis and Speaker Recognition*. Proceedings of ICASSP, San Francisco, USA, vol. 2, pp. II5-II8, March 1992.

[5] ITAKURA, Fumitada. *Minimum Prediction Residual Principle Applied to Speech Recognition*. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-23, number 1, February 1975.

[6] FURUI, S. *Cepstral Analysis Technique for Automatic Speaker Verification*. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, pp. 254-272, April 1981.

[7] HAYKIN, Simon. Adaptive Filter Theory. 3rd. ed. New Jersey: Prentice Hall, 1996.

[8] ALCAIM, Abraham, José Alberto Solewicz, and João Antonio de Morais. *Freqüência de ocorrência dos fonemas e listas de frases foneticamente balanceadas no Português falado no Rio de Janeiro*. Revista da Sociedade Brasileira de Telecomunicações, vol. 7, number 1, December 1992.

[9] DAVIS, Steven B. and Paul Mermelstein. *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences.* IEEE Transactions on Acoustics, Speech, and Signal Processing. vol. ASSP-28, number 4, August 1980.

[10] MARTIN, Alvin and Mark Przybocki. *The NIST 1999 Speaker Recognition Evaluation - An Overview.* Digital Signal Processing, vol. 10, pp. 1-18, 2000.