A FURTHER INVESTIGATION ON AR-VECTOR MODELS FOR TEXT-INDEPENDENT SPEAKER IDENTIFICATION

Ivan MAGRIN-CHAGNOLLEAU *Joachim WILKE* *Frédéric BIMBOT*
Télécom Paris (E.N.S.T.), Dépt. Signal – C.N.R.S., URA 820
46, rue Barrault – 75634 Paris cedex 13 – FRANCE – European Union
email: ivan@sig.enst.fr and bimbot@sig.enst.fr

ABSTRACT

In this paper, we investigate on the role of dynamic information on the performances of AR-vector models for speaker recognition. To this purpose, we design an experimental protocol that destroys the time structure of speech frame sequences, which we compare to a more conventional one, i.e. keeping the natural time order. These results are also compared with those obtained with a (single) Gaussian model. Several measures are systematically investigated in the three cases, and different ways of symmetrisation are tested. We observe that the destruction of the time order can be a factor of improvement for the AR-vector models, and that results obtained with the Gaussian model are merely always better. In most cases, symmetrisation is beneficial.

1. INTRODUCTION

Auto-Regressive (AR) Vector Models have been a significant subject of interest in the field of Speaker Recognition [1] [2] [3] [4] [5] [6] [7]. Whereas the idea of modeling a speaker by an AR-vector model estimated on sequences of speech frames is common to these works, the way to measure the similarity between two speaker models is addressed very differently. Secondly, the use of AR-vector model is often motivated by the belief that such an approach is an efficient way to extract dynamic speaker characteristics, as opposed to static characteristics such as the distribution of speech frame parameters.

In this paper we report on a systematic investigation on similarity measures between AR-vector speaker models obtained as simple combinations of canonical quantities. We also design a protocol in order to examine the role of dynamic information on the performance of the AR-vector approach : we destroy the natural time order of speech frames by shuffling them randomly, and we evaluate the AR-vector approach on these temporally disorganised data. We finally compare both previous approaches to a (single) Gaussian Model [8] [9] [10] [11].

2. DEFINITIONS AND NOTATION

Let $\{\mathbf{x}_t\}_{1\leq t\leq M}$ be a sequence of p-dimensional vectors. Let us define the centered vectors $\mathbf{x}_t^* = \mathbf{x}_t - \bar{\mathbf{x}}$ where $\bar{\mathbf{x}}$ is the mean vector of $\{\mathbf{x}_t\}$.

Let us denote $\mathcal{X}_0$ the covariance matrix of $\{\mathbf{x}_t\}$:

$$\mathcal{X}_0 = \frac{1}{M}\sum_{t=1}^{M}(\mathbf{x}_t - \bar{\mathbf{x}})\cdot(\mathbf{x}_t - \bar{\mathbf{x}})^T = \frac{1}{M}\sum_{t=1}^{M}\mathbf{x}_t^*\cdot\mathbf{x}_t^{*T}$$

We also define as $\mathcal{X}_k$ the lagged covariance matrices :

$$\mathcal{X}_k = \frac{1}{M}\sum_{t=k+1}^{M}\mathbf{x}_t^*\cdot\mathbf{x}_{t-k}^{*T} \text{ with } k = 1, ..., q$$

and the Toeplitz matrix X :

$$X = \begin{bmatrix} \mathcal{X}_0 & \mathcal{X}_1 & ... & \mathcal{X}_q \\ \mathcal{X}_1^T & \mathcal{X}_0 & ... & \mathcal{X}_{q-1} \\ \vdots & \vdots & & \vdots \\ \mathcal{X}_q^T & \mathcal{X}_{q-1}^T & ... & \mathcal{X}_0 \end{bmatrix}$$

A q-th order AR-vector model of sequence $\{\mathbf{x}_t^*\}$ is classically written as :

$$\sum_{i=0}^{q} A_i\cdot\mathbf{x}_{t-i}^* = \mathbf{e}_t \text{ with } A_0 = I_p$$

where $\{A_i\}$ is a set of $q+1$ matrix prediction coefficients, and $\mathbf{e}_t$ is the prediction error vector. $\{A_1, ..., A_q\}$ are obtained by solving the vector Yule-Walker equation [12]. With $A = [A_0 \ ... \ A_q]$, the covariance matrix of the residual of $\{\mathbf{x}_t^*\}$ filtered by A is :

$$E_X^{(A)} = AXA^T$$

Similarily, for a signal $\{\mathbf{y}_t\}_{1\leq t\leq N}$ with model B, we will denote :

$$E_Y^{(B)} = BYB^T$$

If we now consider :

$$\begin{aligned} E_X^{(B)} &= BXB^T \\ E_Y^{(A)} &= AYA^T \end{aligned}$$

these matrices can be interpreted as the covariance matrix of the filtering of $\{\mathbf{x}_t^*\}$ by B, and vice-versa. As A is obtained by minimising $tr(E_X^{(A)})$ and B by minimising $tr(E_Y^{(B)})$, we have $tr(E_X^{(B)}) \geq tr(E_X^{(A)})$ and $tr(E_Y^{(A)}) \geq tr(E_Y^{(B)})$.

Let us finally define $\Gamma_X^{(B/A)}$ and $\Gamma_{Y/X}^{(A)}$ as :

$$\begin{aligned} \Gamma_X^{(B/A)} &= \left(E_X^{(A)}\right)^{-\frac{1}{2}}\cdot E_X^{(B)}\cdot\left(E_X^{(A)}\right)^{-\frac{1}{2}} \\ \Gamma_{Y/X}^{(A)} &= \left(E_X^{(A)}\right)^{-\frac{1}{2}}\cdot E_Y^{(A)}\cdot\left(E_X^{(A)}\right)^{-\frac{1}{2}} \end{aligned}$$

function f	a	$\log a$	g	$\log g$	$a - \log g - 1$	$\log(a/g)$	$a - g$
AR-vector model - spectral frames in their natural time order							
$f_X^{(B/A)} \mid f_Y^{(A/B)}$	16.8 | 8.6	16.8 | 8.6	16.2 | 7.6	16.2 | 7.6	19.1 | 10.8	23.8 | 19.4	22.2 | 17.5
symmetrised	3.5 $^\bullet$	4.1 $^\bullet$	4.1 $^\bullet$	4.1 $^\bullet$	3.2 $^\bullet$	7.9 $^\bullet$	7.3 $^\bullet$
$f_{Y/X}^{(A)} \mid f_{X/Y}^{(B)}$	75.6 | 51.4	75.6 | 51.4	88.3 | 73.0	88.3 | 73.0	15.2 | 34.3	7.6 | 18.7	15.2 | 14.6
symmetrised	6.0 *	4.8 *	12.4 *	4.8 *	5.4 $^\diamond$	7.0 $^\diamond$	6.0 $^\diamond$
AR-vector model - spectral frames in a random time order							
$f_{X'}^{(B'/A')} \mid f_{Y'}^{(A'/B')}$	2.5 | 56.5	2.5 | 56.5	4.1 | 58.1	4.1 | 58.1	2.5 | 56.2	4.1 | 55.9	3.5 | 54.6
symmetrised	3.5 $^\diamond$	3.5 $^\diamond$	5.7 $^\diamond$	5.7 $^\diamond$	2.5 $^\diamond$	4.1 $^\diamond$	4.1 $^\diamond$
$f_{Y'/X'}^{(A')} \mid f_{X'/Y'}^{(B')}$	42.5 | 45.4	42.5 | 45.4	98.1 | 82.9	98.1 | 82.9	**<u>1.3</u>** | 22.9	**<u>1.0</u>** | 6.7	3.2 | 8.9
symmetrised	4.8 *	2.2 *	46.7 *	12.7 *	2.9 $^\diamond$	**<u>1.0</u>** $^\diamond$	**<u>1.6</u>** $^\diamond$
Gaussian model							
$f_{Y_o/X_o}^{(I)} \mid f_{X_o/Y_o}^{(I)}$	37.5 | 47.0	37.5 | 47.0	98.4 | 98.4	98.4 | 98.4	**<u>0.6</u>** | 7.9	**<u>0.6</u>** | 3.2	2.9 | 6.4
symmetrised	3.8 *	**<u>1.3</u>** *	97.1 *	99.4 *	**<u>1.0</u>** $^\diamond$	**<u>0.6</u>** $^\diamond$	**<u>1.0</u>** $^\diamond$

Table 1. TIMIT - Speaker identification error rates

where $E^{\frac{1}{2}}$ is the symmetric square root matrix of E.
The first matrix can be interpreted as the covariance matrix of $\{\mathbf{x}_t^*\}$ filtered by B relative to the one of $\{\mathbf{x}_t^*\}$ filtered by A, and the second one as the covariance matrix of $\{\mathbf{y}_t^*\}$ filtered by A relative to the one of $\{\mathbf{x}_t^*\}$ filtered by A.

3. SPEAKER MODELS

The purpose of this paper is to investigate on different ways of using an AR-vector model for speaker identification. A speaker is characterised by a second-order AR-vector model ($q = 2$) estimated on some speech material training. The matrix prediction coefficients $\{A_1, A_2\}$ are obtained by solving the vector Yule-Walker equation in the case $q = 2$:

$$[A_1\ A_2] \cdot \begin{bmatrix} \mathcal{X}_0 & \mathcal{X}_1 \\ \mathcal{X}_1^T & \mathcal{X}_0 \end{bmatrix} = -\left[\mathcal{X}_1^T\ \mathcal{X}_2^T\right]$$

- A first model is a 2nd-order AR-vector model trained on speech frames presented in their natural time order. Therefore, the model of $\mathcal{X}$ is $\{A, X\}$.
- A second model is a 2nd-order AR-vector model trained on the same speech frames as previously, but presented in a random time order. Each speaker $\mathcal{X}$ is characterised by $\{A', X'\}$ which are obtained in the same way as $\{A, X\}$, after speech frames have been randomly shuffled.

Gaussian speaker model is also tested as a reference model. In this second framework, a speaker $\mathcal{X}$ is represented by the covariance matrix $\mathcal{X}_0$. It is equivalent to a 0th-order AR-vector model, i.e. $A = [A_0] = I_p$ and $X = [\mathcal{X}_0]$, which we will denote as $\{I, X_0\}$.

4. SIMILARITY MEASURES

We consider now 2 speakers $\mathcal{X}$ and $\mathcal{Y}$, and we present a general formalism for expressing similarity measures between their AR-vector models.

Two families of similarity measures are investigated :

$$\begin{aligned} f_X^{(B/A)}(\mathcal{X}, \mathcal{Y}) &= f\left(\Gamma_X^{(B/A)}\right) \\ f_{Y/X}^{(A)}\ (\mathcal{X}, \mathcal{Y}) &= f\left(\Gamma_{Y/X}^{(A)}\right) \end{aligned}$$

The first family can be interpreted as a measure between two models (A and B), via their influence on the same vector signal (X). This family of measures (which we will refer to as VI), generalises the Itakura measure to the vector case [13]. Examples of such measures are proposed in [4] and [6]. On the opposite, the second family can be viewed as a measure between two signals (X and Y) filtered by a common model (A). Some of the IS measures proposed in [3] [5] belong to this family. Note also that setting $\{A, X\}$ = $\{I, X_0\}$ allows to construct a similar family of measures for the Gaussian model.
The function f is chosen equal to a combination of the following canonical quantities :

$$\begin{aligned} a(\Gamma) &= \tfrac{1}{p}\, tr(\Gamma) \\ g(\Gamma) &= [det(\Gamma)]^{\frac{1}{p}} \end{aligned}$$

It can be shown that a and g are positive and that $a \geq g$. Moreover these quantities can be computed very efficiently [11]. The composed functions $a - \log g - 1$ and $\log\ (a/g)$ are respectively the Maximum-Likelihood measure [9] and the Arithmetic-Geometric Sphericity measure [8].
As these measures are not symmetric, different symmetrisations can be applied on the original measures. Given $f_X^{(B/A)}$ and $f_Y^{(A/B)}$, we define :

$$\begin{aligned} f_X^{(B/A)^\star} &= \frac{1}{2}\, f_X^{(B/A)} + \frac{1}{2}\, f_Y^{(A/B)} \\ f_X^{(B/A)^\diamond} &= \frac{\bar{M}}{\bar{M}+\bar{N}}\, f_X^{(B/A)} + \frac{\bar{N}}{\bar{M}+\bar{N}}\, f_Y^{(A/B)} \\ f_X^{(B/A)^\bullet} &= \frac{\bar{N}}{\bar{M}+\bar{N}}\, f_X^{(B/A)} + \frac{\bar{M}}{\bar{M}+\bar{N}}\, f_Y^{(A/B)} \end{aligned}$$

function f	a	$\log a$	g	$\log g$	$a - \log g - 1$	$\log(a/g)$	$a - g$
AR-vector model - spectral frames in their natural time order							
$f_X^{(B/A)} \mid f_Y^{(A/B)}$	38.7 | 30.2	38.7 | 30.2	37.1 | 29.5	37.1 | 29.5	42.5 | 35.2	51.1 | 50.8	49.5 | 49.5
symmetrised	24.8 $\bullet$	25.1 $\bullet$	24.8 $\bullet$	24.4 $\bullet$	26.3 $\bullet$	35.6 $\bullet$	33.3 $\bullet$
$f_{Y/X}^{(A)} \mid f_{X/Y}^{(B)}$	93.3 | 86.0	93.3 | 86.0	96.5 | 94.6	96.5 | 94.6	44.1 | 69.8	41.6 | 39.1	49.2 | 39.1
symmetrised	23.5 *	21.3 *	32.4 *	25.4 *	24.4 $^\diamond$	34.6 $^\diamond$	33.0 $^\diamond$
AR-vector model - spectral frames in a random time order							
$f_{X'}^{(B'/A')} \mid f_{Y'}^{(A'/B')}$	35.9 | 82.2	35.9 | 82.2	36.8 | 81.3	36.8 | 81.3	32.4 | 83.5	34.6 | 82.2	34.3 | 81.6
symmetrised	39.1 $^\diamond$	39.1 $^\diamond$	40.0 $^\diamond$	40.0 $^\diamond$	34.3 $^\diamond$	33.3 $^\diamond$	33.3 $^\diamond$
$f_{Y'/X'}^{(A')} \mid f_{X'/Y'}^{(B')}$	78.7 | 71.4	78.7 | 71.4	98.4 | 93.7	98.4 | 93.7	15.9 | 43.8	**13.3** | 21.6	20.3 | 27.3
symmetrisation	21.9 *	**14.6** *	69.8 *	52.4 *	**14.0** $^\diamond$	**13.3** $^\diamond$	**14.3** $^\diamond$
Gaussian model							
$f_{Y_o/X_o}^{(I)} \mid f_{X_o/Y_o}^{(I)}$	77.1 | 71.8	77.1 | 71.8	98.4 | 98.4	98.4 | 98.4	**14.6** | 27.3	**12.7** | 17.1	20.3 | 21.3
symmetrised	15.6 *	**11.8** *	97.8 *	98.4 *	**12.7** $^\diamond$	**12.4** $^\diamond$	**14.3** $^\diamond$

Table 2. FTIMIT - Speaker identification error rates

$\bar{M}$ is the average number of frames for the training sentences across all speakers, and $\bar{N}$ is the average number of frames for the test sentences. The same symmetrisations are applied to $f_{Y/X}^{(A)}$ and $f_{X/Y}^{(B)}$.

5. DATABASE AND SIGNAL ANALYSIS

We use the first 63 speakers of TIMIT [14] and NTIMIT [15] for our experiments (19 females and 44 males)[1]. Each of them has read 10 sentences. The signal is sampled at 16 kHz, on 16 bits, on a linear amplitude scale. NTIMIT is a telephone-channel version of TIMIT.

Each sentence is analysed as follows : for each speech token, the speech signal is kept in its integrality; it is decomposed into frames of 31.5 ms at a frame rate of 10 ms, with no pre-emphasis. A Hamming window is applied to each frame. Then the module of a 504 point Fourier Transform is computed, from which 24 Mel-scale triangular filter bank coefficients are extracted. The spectral vectors $\{\mathbf{x}_t\}$ (of dimension $p = 24$) are formed from the logarithm of each filter output. These analysis conditions are identical to those used in [11].

For the TIMIT database, all 24 coefficients of $\{\mathbf{x}_t\}$ are kept. For NTIMIT, 24-dimensional vectors are also extracted, but we keep only the first 17 coefficients, which corresponds to the telephone bandwidth. Experiments are also made on "FTIMIT", obtained by taking the 17 first coefficients of the vectors $\{\mathbf{x}_t\}$ extracted from TIMIT.

6. EXPERIMENTS

A common training/test protocol is used for all the experiments. It is described in detail in [11] (as protocol "long-short"). Training material consists of 5 sentences (i.e $\approx$ 14.4 s) which are concatenated into a single reference per speaker. Tests are carried out on 5×1 sentence per speaker (i.e $\approx$ 3.2 s per sentence) which are tested separately. The total number of independent tests is therefore $63 \times 5 = 315$. The decision rule is the 1-nearest neighbour.

Results of the experiments are given by database (Tables 1 2 and 3). Performances are reported in terms of closed-set speaker identification error rates on the test set for the canonical measures and various combined measures in their asymmetric and their best symmetric form. For the symmetrised measures, a superscript indicates to which symmetrisation ($\star$, $\diamond$ or $\bullet$) does the result correspond.

7. DISCUSSION

The following observations can be made :

- Symmetrisation is generally a factor of improvement. However, the appropriate symmetrisation is difficult to predict. It depends on the type of asymmetric measure, and whether the data are in a natural or in a random time order.
- For each database (TIMIT, FTIMIT and NTIMIT), we have underlined the best 10 (or 11) measures. They are (almost) the same ones for all 3 databases. The best one is always obtained with the Gaussian Model.
- With spectral frames in their natural order, VI measures globally outperform IS measures in canonical forms, but the trend is inverted with composed forms.
- With spectral frames in a random order, symmetric composed IS measures outperform all other AR-vector measures, in spite of the loss of the dynamic spectral characteristics.

8. CONCLUSION

In our experiments, we did not succeed in obtaining better speaker identification results with an AR-vector model based measure than with a single Gaussian model classifier.

[1] More precisely, we have kept all female and male speakers of "train/dr1" and "test/dr1", the first female speaker of "train/dr2", and the first 13 male speakers of "train/dr2".

function f	a	$\log a$	g	$\log g$	$a - \log g - 1$	$\log (a/g)$	$a - g$
AR-vector model - spectral frames in their natural time order							
$f_X^{(B/A)} \mid f_Y^{(A/B)}$	71.8 | 54.6	71.8 | 54.6	67.3 | 54.3	67.3 | 54.3	78.1 | 58.4	83.8 | 69.5	82.9 | 67.9
symmetrised	51.8 •	52.1 •	50.5 •	50.2 •	57.5 •	66.0 •	65.1 •
$f_{Y/X}^{(A)} \mid f_{X/Y}^{(B)}$	96.8 | 92.4	96.8 | 92.4	97.1 | 95.6	97.1 | 95.6	67.3 | 88.9	66.0 | 78.7	75.2 | 76.8
symmetrised	61.9 ⋆	56.5 ⋆	68.3 ⋆	53.0 ⋆	59.7 ◇	63.2 ◇	66.4 ◇
AR-vector model - spectral frames in a random time order							
$f_{X'}^{(B'/A')} \mid f_{Y'}^{(A'/B')}$	64.4 | 92.1	64.1 | 92.1	65.4 | 91.8	65.4 | 91.8	61.9 | 92.4	64.8 | 93.3	64.4 | 93.0
symmetrised	65.4 ◇	65.1 ◇	67.9 ◇	68.3 ◇	62.2 ◇	64.4 ◇	64.1 ◇
$f_{Y'/X'}^{(A')} \mid f_{X'/Y'}^{(B')}$	94.0 | 94.3	94.0 | 94.3	98.4 | 97.5	98.4 | 97.5	**47.0** | 86.4	**46.0** | 63.2	56.8 | 77.1
symmetrisation	61.9 ⋆	52.4 ⋆	88.3 ⋆	72.4 ⋆	50.2 ◇	**44.1** ◇	**48.6** ◇
Gaussian model							
$f_{Y_o/X_o}^{(I)} \mid f_{X_o/Y_o}^{(I)}$	93.0 | 94.6	93.0 | 94.6	98.4 | 98.4	98.4 | 98.4	**44.1** | 75.9	**42.5** | 59.7	56.2 | 73.3
symmetrised	58.1 ⋆	**49.8** ⋆	97.8 ⋆	98.4 ⋆	**47.6** ◇	**44.1** ◇	**49.2** ◇

Table 3. NTIMIT - Speaker identification error rates

This observation is in contradiction with results reported in [7], but this divergence may be due to different signal pre-processing and analysis.

Moreover, we globally obtained better performances with the AR-vector model on spectral frames in a random time order rather than when we kept the natural time order. Therefore, the role of dynamic speaker characteristics in the success of the AR-vector model can be questioned, as our results suggest that AR-vector models tend to extract indirectly speaker characteristics of a static nature.

Finally, the influence of symmetrisation can be crucial, but its theoretical basis remains to be understood.

REFERENCES

[1] Yves Grenier. Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur. In *XIèmes Journées d'Etude sur la Parole*, pages 163–171, May 1980. Strasbourg, France.

[2] T. Artières, Y. Bennani, P. Gallinari, and C. Montacié. Connectionnist and conventional models for text-free talker identification tasks. In *Proceedings of NEURONIMES 91*, 1991. Nîmes, France.

[3] C. Montacié, P. Deléglise, F. Bimbot, and M.-J. Caraty. Cinematic techniques for speech processing: temporal decomposition and multivariate linear prediction. In *Proceedings of ICASSP 92*, volume 1, pages 153–156, March 1992. San Francisco, United-States.

[4] F. Bimbot, L. Mathan, A. de Lima, and G. Chollet. Standard and target-driven AR-vector models for speech analysis and speaker recognition. In *Proceedings of ICASSP 92*, volume 2, pages II.5–II.8, March 1992. San Francisco, United-States.

[5] Claude Montacié and Jean-Luc Le Floch. AR-vector models for free-text speaker recognition. In *Proceedings of ICSLP 92*, volume 1, pages 611–614, October 1992. Banff, Canada.

[6] Chintana Griffin, Tomoko Matsui, and Sadoaki Furui. Distance measures for text-independent speaker recognition based on MAR model. In *Proceedings of ICASSP 94*, volume 1, pages 309–312, April 1994. Adelaïde, Australia.

[7] J.-L. Le Floch, C. Montacié, and M.-J. Caraty. Speaker recognition experiments on the NTIMIT database. In *Proceedings of EUROSPEECH 95*, volume 1, pages 379–382, September 1995. Madrid, Spain.

[8] Yves Grenier. *Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonémique*. PhD thesis, ENST, 1977.

[9] Herbert Gish, Michael Krasner, William Russell, and Jared Wolf. Methods and experiments for text-independent speaker recognition over telephone channels. In *Proceedings of ICASSP 86*, volume 2, pages 865–868, April 1986. Tokyo, Japan.

[10] Frédéric Bimbot and Luc Mathan. Text-free speaker recognition using an arithmetic-harmonic sphericity measure. In *Proceedings of EUROSPEECH 93*, volume 1, pages 169–172, September 1993. Berlin, Germany.

[11] Frédéric Bimbot, Ivan Magrin-Chagnolleau, and Luc Mathan. Second-order statistical measures for text-independent speaker identification. *Speech Communication*, 17(1–2):177–192, August 1995.

[12] P. Whittle. On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika*, 50(1–2):129–134, 1963.

[13] Fumitada Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):67–72, February 1975.

[14] William M. Fisher, George R. Doddington, and Kathleen M. Goudie-Marshall. The DARPA speech recognition research database : specifications and status. In *Proceedings of the DARPA workshop on speech recognition*, pages 93–99, February 1986.

[15] Charles Jankowski, Ashok Kalyanswamy, Sara Basson, and Judith Spitz. NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. *Proceedings of ICASSP 90*, April 1990. New Mexico, United-States.