БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ . BULGARIAN ACADEMY OF SCIENCES

КИБЕРНЕТИКА И ИНФОРМАЦИОННИ ТЕХНОЛОГИИ · Том **3**, № 1
CYBERNETICS AND INFORMATION TECHNOLOGIES · Volume **3**, No 1

София . 2003 . Sofia

An Evaluation of DTW, AA and ARVM for Fixed-Text Speaker Identification

Atanas Ouzounov

Institute of Information Technologies, 1113 Sofia, E-mail: atanas@iinf.bas.bg

Abstract: *Three different methodologies for automatic speaker identification have been evaluated in the paper, namely the well known Dynamic Time Warping (DTW), the Auto-Regressive Vector Models (ARVM) and an Algebraic Approach (AA). The aim of our study is to examine the effectiveness of these approaches in the fixed-text speaker identification task with short phrases in Bulgarian language collected over noisy telephone channels. Furthermore, two well-known speech features, namely the Linear Predictive Coding derived Cepstrum (LPCC) and the Mel-Frequency Cepstral Coefficients (MFCC) were evaluated. As experimental results shown the joint work of the ARVM and the MFCC outperforms the all others approaches used in this study.*

Keywords: *speaker identification, mel-frequency cepstrum, linear predictive coding cepstrum, algebraic approach.*

1. Introduction

Automatic speaker identification comprises two groups of methods: for text-dependent (fixed-text) and for text-independent (free-text) identification. In the case of fixed-text identification, it is required that the speaker uses the same phrase or sentence in both training and recognition modes, whereas in text-independent there is no such constraint. The most popular algorithms for text-dependent identification are Dynamic Time Warping (DTW) and Hidden Markov's Models (HMM). The HMM achieve better recognition rate compared to the DTW, but at the cost of higher number of computations in the training mode [3].

The Algebraic Approaches (AAs) can be considered as one traditional strategy in the speaker identification and they are usually used in text-independent speaker identification tasks [1]. They are based on an estimation of the covariance matrix of speech data and it is known that the accuracy of this estimation depends on the amount

of used data. Usually in text-independent speaker identification task, the amount of data is few times more than in the fixed-text recognition case. For short phrases with length of about 1–2 s (typical phrase length for fixed-text identification tasks) the covariance matrix estimation would be poor, especially if we process noisy telephone speech.

In fact, the covariance matrix used in algebraic approaches is a static model of the speaker voice. An important issue in the field of speaker identification is an estimation of the dynamics of the speaker voice. This can be done by the Auto-Regressive (AR) vector modelling of the speech data [5]. Standard AR-vector modelling is a generalization of the vector case of the well-known scalar auto-regressive modelling technique. The parameters of Auto-Regressive Vector Models (ARVM) can be estimated by a vector version of Levinson's algorithm. In last decade the ARVM are usually used in text-independent speaker recognition tasks [5].

The smaller number of computations for references creation (i.e. short training time) distinguishes the mentioned above approaches (DTW, AA and ARVM) by HMM, neural networks and vector quantization (VQ) [3]. It is interesting to note that in [2] the author compares the DTW and the VQ approaches in a text-dependent speaker verification task with telephone speech data and concludes that the DTW overwhelms the VQ in almost all tests.

The short training time is an important feature if we want to develop an automated fixed-text speaker identification system running on a PC without using a dedicated hardware. This feature motivated us to compare the performance of AA and ARVM to the performance of the famous method in the fixed-text speaker recognition area – the DTW. It is worth to note again that the algebraic approaches and the auto-regressive vector models usually are not used in the text-dependent speaker identification tasks.

It is known that these three approaches use three different strategies in recognition process. While the DTW is based on the vector sequences time alignment then the AA uses parameters obtained from the covariance matrices of these sequences. On the other hand, the auto-regressive vector model allows for the form of the features trajectories of the analyzed speech data. It is unclear which one of these strategies will be more effective in our case when the data is noisy telephone speech and its length is few seconds.

It is known that recognition performance depends on the chosen combination between features and classification rule. Wherefore we evaluated here two parametric presentations – the Linear Predictive Coding derived Cepstrum (LPCC) and Mel-Frequency Cepstral Coefficients (MFCC) [7].

In our experiments we study the effects of the length of the speech data on the performance of selected approaches in the fixed-text speaker identification task with short phrases in Bulgarian language collected over noisy telephone channels.

2. Speech features

In our study we chose as speech features two well-known parametric presentations widely used in speech and speaker recognition experiments namely, LPCC and MFCC [7].

To calculate the LPCC, the Linear Predictive Coding (LPC) coefficients must be first calculated. Then the cepstral coefficients $c_{\mathrm{LPC}}(k)$ can be computed by the following recursion [7, 8]:

$$c_{\mathrm{LPC}}(k) = \begin{cases} -a(k) - \sum_{i=1}^{k-1}\left(1 - \frac{i}{k}\right)a(i)c_{\mathrm{LPC}}(k-i), & 1 \le k \le p, \\ -\sum_{i=1}^{k-1}\left(1 - \frac{i}{k}\right)a(i)c_{\mathrm{LPC}}(k-i), & k > p, \end{cases} \tag{1}$$

where $a(k)$, $k = 1, ..., p$ are LPC coefficients and p is the model order.

In order to compute the MFCC of a speech segment we calculate the power spectrum of the speech frame. Then, we pass the power spectrum through each filter of the filter bank, calculating the output power of the filters. The mel-frequency cepstral coefficients $c_{\mathrm{MEL}}(m)$ are computed by the following formula [8]:

$$c_{\mathrm{MEL}}(m) = \sum_{k=1}^{K} \log(S(k)) \cos(m(k-0.5)\frac{\pi}{K}), \tag{2}$$

where K is the number of triangular bandpass mel-frequency scale filters, $S(k)$ is the output power of k-th mel filter and $m = 1,\ldots, M$ is the cepstral coefficients index, $c(0)$ is not used.

3. Speaker modelling

3.1. Dynamic Time Warping [6]

In our work, we apply the modified DTW algorithm, called the normalize-wrap method. In this algorithm, we use the length normalization on both the reference and test pattern before performing the actual DTW algorithm. In the DTW, we implement the relaxed endpoints constraints, Itakura's form of local constraints and Euclidean cepstral distance as local distance.

We place the reference along the Y-axis and set the path width at 300 ms. The speaker's reference is obtained by averaging (after dynamic time warping alignment) of his training utterances. Fig.1 shows the block diagram of the DTW training and identification procedures.

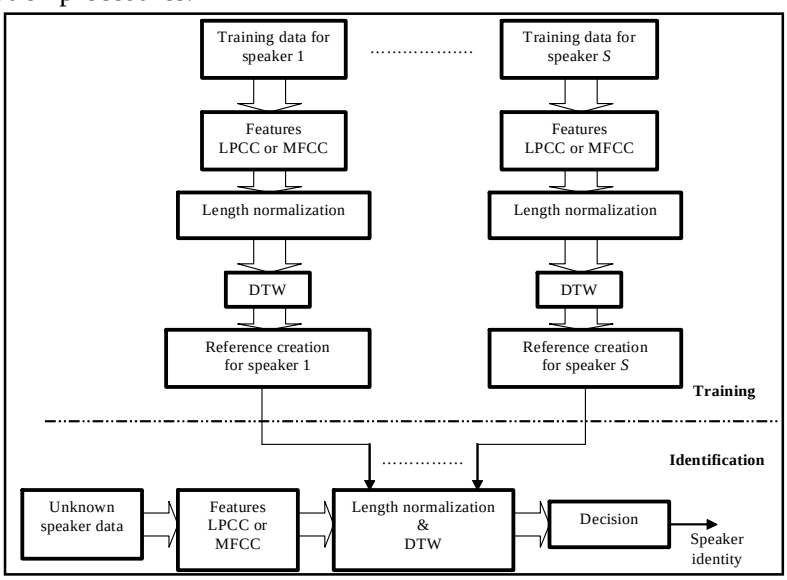

Fig.1. Block diagram of the DTW training and identification procedures

3.2. Algebraic approach [1]

From the sampled speech signal we compute parametric presentation – sequence of N vectors in *p*-dimensional space. Let $\{X_n\}$, $n = 1, ..., N$, is this vector sequence. The covariance matrix C_X of $\{X_n\}$ is obtained as [1]

$$C_X = \frac{1}{N}\sum_{n=1}^{N} X_n X_n^{\mathrm{T}} - m_x m_x^{\mathrm{T}}, \tag{3}$$

$$m_X = \frac{1}{N}\sum_{n=1}^{N} X_n, \tag{4}$$

where m_X is the mean vector of $\{X_n\}$ and T is matrix transposition. The covariance matrix of a test sequence $\{Y_l\}$, $l = 1, ..., L$, is denoted as C_Y.

There are measures which can be considered as different estimations of the similarity of two covariance matrices [1,5]. They belong to the distances that could be defined using only the eigenvalues of the product $C_Y C_X^{-1}$. Let the *p*-ordered eigenvalues of the matrix $C_Y C_X^{-1}$ are denoted as $\{\lambda_i\}$ and three particular functions of $\{\lambda_i\}$ are defined [1]:

$$A(\lambda_1,...,\lambda_p) = \frac{1}{p}\sum_{i=1}^{p}\lambda_i; \tag{5}$$

$$G(\lambda_1,...,\lambda_p) = \sqrt[p]{\prod_{i=1}^{p}\lambda_i}; \tag{6}$$

$$H(\lambda_1,...,\lambda_p) = p\left(\sum_{i=1}^{p}\frac{1}{\lambda_i}\right)^{-1}. \tag{7}$$

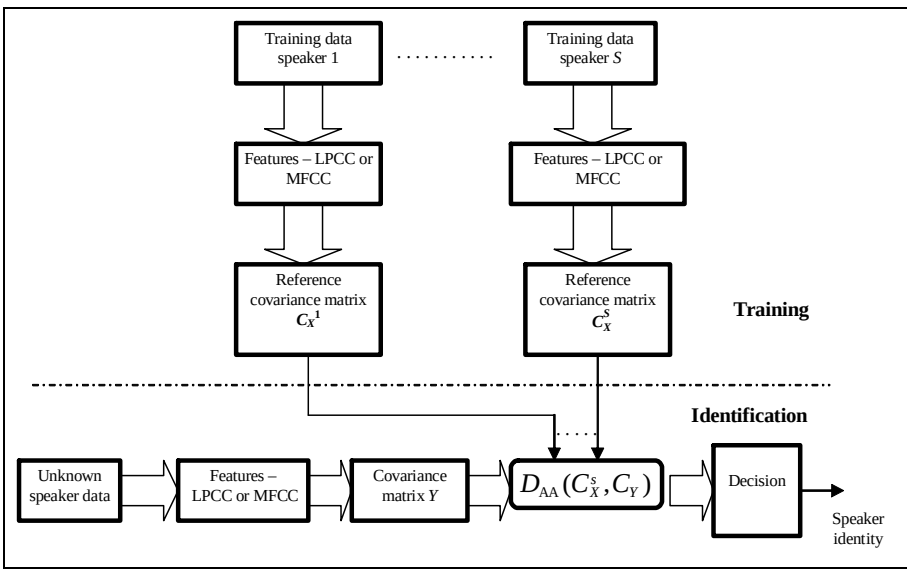

Fig. 2. Block diagram of the AA training and identification method

Functions A,G and H are respectively arithmetic, geometric and harmonic means of the eigenvalues $\{\lambda_i\}$. The swapping of the matrices C_X and C_Y leads to transformation of A into H^{-1}, G into G^{-1} and H into A^{-1}.

Various distance measures were constructed based on these mean values [1, 4]. Here we will use a distance proposed in [4]. This distance $D_{\mathrm{AA}}(C_X, C_Y)$ is non-symmetric and it is in the form [4]:

$$D_{\mathrm{AA}}(C_X, C_Y) = \log\left(\frac{A^2}{GH}\right). \tag{8}$$

The speaker's reference consists of his reference covariance matrix. For a particular speaker this matrix is obtained by averaging of the covariance matrices of his training utterances. In testing mode the distance $D_{\mathrm{AA}}(C_X^s, C_Y)$ in (8) is calculated between covariance matrix C_Y obtained from input speech sequence from unknown speaker and reference matrices C_X^s of all speakers $s = 1, ..., S$. Fig.2 shows the block diagram of the procedures for training and identification in the AA method.

3.3. Auto-regressive vector modeling [5]

The AR-vector models are used here to describe the parametric vectors' trajectories of the analyzed speech data. Let $X = \{X_n\}$, $n = 1, ..., N$, be the vector sequence in p-dimensional space. Each vector can be described by an auto-regressive vector model of order q in the form similar to the scalar case

$$\sum_{k=0}^{q} A_k X_{n-k} = e_n, \tag{9}$$

where, $\{A_k\}$, $k = 0, ..., q$, are prediction coefficient matrices with size $p \times p$ and $\{e_n\}$, $n = 1, ..., N$, is the prediction error vectors with size p. The matrix coefficients of the model can be estimated by the vector version of the Levinson's algorithm. The criterion to minimize is the trace of covariance matrix of $\{e_n\}$.

As a similarity measure between two AR-vector models (reference A and test B) the measure of their influence on the same vector sequence (test sequence Y) is used. This measure $D_\Gamma(A, B)$ is non-symmetric and it is based on determinant of the matrix $\Gamma_Y^{(A/B)}$ [5]. The matrix is

$$\Gamma_Y^{A/B} = \left(E_Y^B\right)^{-1/2} E_Y^A \left(E_Y^B\right)^{-1/2}, \tag{10}$$

where $A = \{A_k\}$, $k = 0, ..., q$, is the vector model of size q obtained from the reference sequence X;

$B = \{B_k\}$, $k = 0, ..., q$, is the vector model of size q obtained from the test sequence Y;

E_Y^B is the covariance matrix of the residual vector sequence of Y filtered by model B;

E_Y^A is the covariance matrix of the residual vector sequence of Y filtered by model A;

The non-symmetric measure used is:

$$D_\Gamma(A, B) = [\det(\Gamma_Y^{A/B})]^{1/p}. \tag{11}$$

The speech data is processed by 2nd order AR-vector model ($q = 2$). The speaker's reference consists of the matrix prediction coefficients obtained from his common block-autocorrelation matrix.

For a particular speaker this matrix is estimated by averaging of the block-autocorrelation matrices of the vector sequences of his training utterances.

In testing mode the distance $D_\Gamma(A, B)$ in (11) is calculated between the AR-vector model B obtained from input speech sequence Y and the reference AR-vector models A^s of all speakers $s = 1,..., S$. Fig.3 shows the block diagram of the training and identification procedures in the ARVM method.

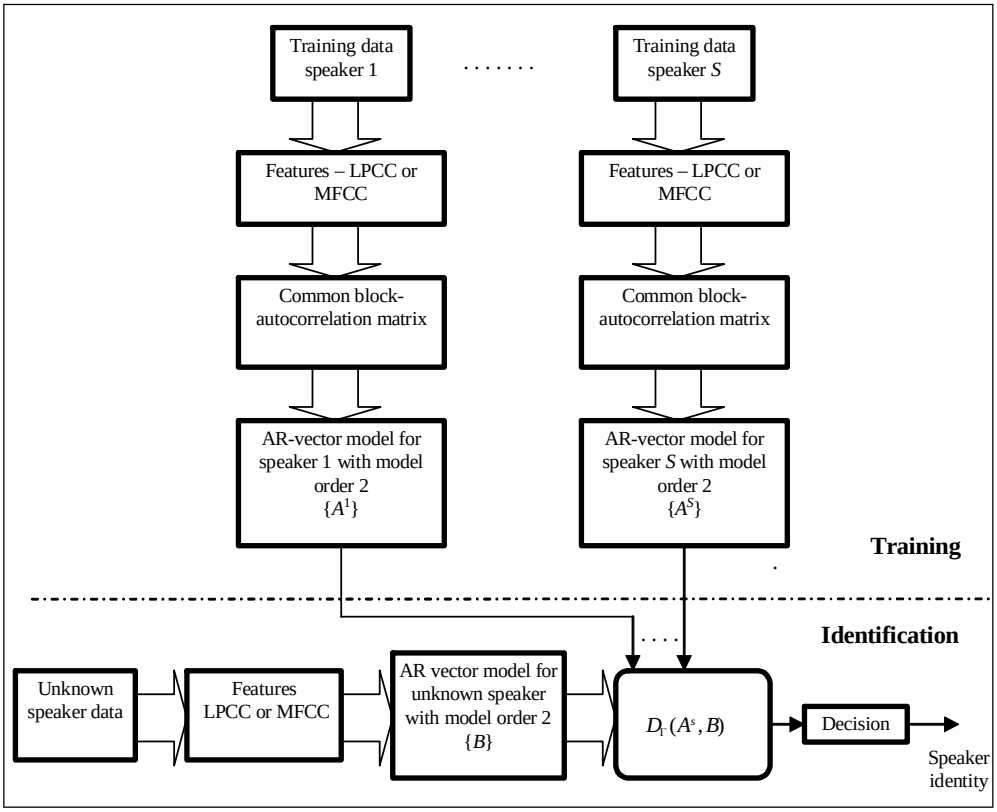

Fig.3. Block diagram of training and identification procedures in the ARVM method

4. Experiments

The analysed phrase in Bulgarian language is in length of about 2 s. The records were made over dialled-up telephone (analogue) lines in the city of Sofia. The speech signal is digitized at 8 kHz on 16 bits, after low-pass filtering at 4 kHz. Preemphasis is not applied. Hamming windowing frames of 32 ms are used, with frame rate of 8 ms. A 14th order autocorrelation analysis is carried out. Each frame is then converted into 14th order LPC derived cepstral vector. The MFCC are 14 and they are calculated by using the 24 triangular mel-frequency filters.

Since the used speech database consists of short phrases with length up to 2 s, we did not apply any frames selection (e.g. voiced/unvoiced or speech/non-speech). We always used all frames in the phrase. Only manual phrase endpoints detection is performed to avoid the processing of non-speech parts of the recorded signal.

It is known that the different channel characteristics during training and testing sessions can seriously degrade the performance of speaker identification. Therefore, channel compensation technique by cepstral mean subtraction is applied [3].

The used database comprises speech material from 12 male speakers – about 20 phrase repetitions per speaker. Every repetition is obtained from different telephone call. First 8 repetitions form the speaker's training data. During the training procedure we do not make any selection among the utterances intended for reference creation, i.e., here we use the multistyle training. That means we can form the reference from data consists of a clear utterance together with a very noisy one.

The effectiveness of all algorithms for closed-set fixed-text speaker identification is estimated as a function of the length of training data. Experimental results are shown in Table 1. The number of tests is 134 and the identification error is averaged over 12 speakers. Since there are different number of test utterances per speaker we first compute the individual errors and then average them to produce the final errors shown in Table 1. The length of test (one utterance) is about 2 s. The test utterances are not included in the training data. The length of training data 6, 10 and 16 seconds corresponds to 3, 5 and 8 utterances used for training. The identification is based on minimum distance rule. No additional threshold is used.

Table 1. Identification error in percentage

Training data in s	Test (approx. 2 s)					
	DTW		AA		ARVM	
	LPCC	MFCC	LPCC	MFCC	LPCC	MFCC
≈6	39.20	39.44	37.33	27.61	33.33	24.99
≈10	28.73	25.21	28.36	19.39	20.99	20.24
≈16	20.40	19.87	19.33	19.97	16.13	15.49

5. Conclusions

The results in Table 1 shown that the joint work of ARVM and MFCC outperforms the all others approaches used in our study. This fact surprised us because the ARVM is intended for text-independent speaker recognition. This technique is based on an estimation of block-autocorrelation matrix that depends on the amount of data. We supposed that for short training data the performance of the ARVM technique would be worse than DTW one, but the experiments yielded the opposite result. About the covariance matrix based approach (AA) we can say that it took a middle place in our final results arrangement. Its results were closer to the DTW's results than to ARVM ones.

To prove that the modelling of the feature trajectories is more effective strategy in the recognition process (even for short noisy speech data) than the direct vector sequences time alignment or covariance matrices comparison.

These experimental results suggest that for fixed-text speaker identification tasks with short training time we can prefer the ARVM–MFCC technique to both AA and DTW. We plan in our forthcoming work to examine the joint work of the ARVM and others more robust features intended especially for noisy telephone speech data.

References

1. B i m b o t, F., I. M a r g i n, L. M a t h a n. Second-order statistical measures for text-independent speaker identification. – Speech Communication, **17**, 1995, No 1-2, 177-192.
2. F a l c o n e, M. Evaluation of two algorithms for text-dependent speaker verification on real telephone speech for Italian language. – ESPRIT Project 6819 SAM-A, 1993, 1-7.
3. F u r u i, S. An overview of speaker recognition technology. – In: ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, 1994, 1-9.
4. J o h n s o n, S. Speaker tracking. Mphil thesis, University of Cambridge, 1997.
5. M a g r i n - C h a g n o l l e a u, I., J. W i l k e, F. B i m b o t. A further investigation on AR-vector models for text-independent speaker identification. – In: Proceedings of the ICASSP 96, 1996, 401-404.
6. M y e r s, C., L. R a b i n e r, A. R o s e n b e r g. Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition. – IEEE Transactions on ASSP, **ASSP**-**28**, 1980, No 6, 623-635.
7. P i c o n e, J. Signal Modeling Techniques in Speech Recognition. – In: Proceedings of the IEEE, **81**, 1993, No 9, 1215-1247.
8. R a b i n e r, L., B.- H. J u a n g. Fundamentals of Speech Recognition. Prentice Hall, 1993.

Сравнителен анализ на ДП, АРВМ и АМ при зависима от текста идентификация на диктори

Атанас Узунов

Институт по информационни технологии – 1113, София, E-mail: atanas@iinf.bas.bg

(Р е з ю м е)

В статията е извършен сравнителен експериментален анализ на три метода за разпознаване на диктори, а именно динамично програмиране (ДП), авторегресионни векторни модели (АРВМ) и един алгебричен метод (АМ). Целта на този анализ е да се установи ефективността на разгледаните методи, при зависима от текста идентификация на диктори, за кратки фрази на български език, записани по телефонен канал. Като параметрични представяния са използвани две кепстрални представяния: кепстърът, получен чрез метода на линейно предсказване, и мел-кепстърът. Експерименталните резултати показаха, че съвместната работа на мел-кепстъра и авто-регресионните векторни модели превъзхожда от гледна точка на точността на разпознаване останалите методи, включени в настоящето изследване.