Performance Evaluation of an Automatic Forensic Speaker Recognition System based on GMM

Francesco Beritelli and Andrea Spadaccini
Dipartimento di Ingegneria Informatica e delle Telecomunicazioni
Facoltà di Ingegneria - Università di Catania
Viale A. Doria 6 - Catania - Italy
E-Mail: {beritelli,spadaccini}@diit.unict.it

***Abstract*—This paper presents a performance evaluation of a speech biometry system based on the statistical models GMM (Gaussian Mixture Models). In particular, the paper underlines the robustness to the degradation of various natural noises, and their impact on the system. Finally, the impact of the duration to both training and test sequences is highlighted. Results show that the noise can have the impact on the degradation of the performance (see EER values) which vary from 100 % to 300 % on the basis of the type of noise which depends on only one of two compared sequences. The duration of the sequences is a very important parameter, mostly for training phase, for which it is necessary to have at least 25 seconds long talk.**

***Index Terms*—speaker recognition; SNR estimation; voice/noise detection; forensic biometry**

I. Introduction

One of the most important aims of the speech biometrics in a forensic setting regards the study of new methods contributing to automatic speaker identification. The most important points examine automatic system performance in the presence of short and/or bad quality phonic signals. Speech signal quality is of fundamental importance for accurate speaker identification [1]. The reliability of a speech biometry system is known to depend on the amount of material available, in particular on the number of vowels present in the sequence being analysed, and the quality of the signal [2]. The former affects the resolution power of the system, while the latter impacts the correct estimation of biometric indexes. In automatic or semi-automatic speaker recognition, background noise is one of the main causes of alteration of the acoustic indexes used in the biometric identification/verification phase [3]. Therefore, background noise is one of the main causes of a performance degradation of a biometry system. In general, the behaviour of a speech processing system depends on the level of background noise and also on the type of noise, which may be of various kinds (car, bus, office, restaurant, street, stadium, etc.), presenting in each case different statistical and spectral characteristics. This paper presents vocal biometry system performance based on the statistical models GMM (Gaussian Mixture Models). In particular, the paper underlines both the performance using "clean" database and the robustness to the degradation of various natural noises, and their impact on the system. Finally, the impact of the duration to both training and test sequences is highlighted.

II. Gaussian Mixture Models for Speaker Identity Recognition

In this section we present an overview of a well-known statistical method for speaker recognition: Gaussian Mixture Models (GMMs).

We can restate the problem of determiming whether a given input speech signal s belongs to a given identity I as an hypothesis test between two hypotheses:

$$H_0 : s \text{ belongs to } I$$
$$H_1 : s \text{ does not belong to } I$$

This decision can be taken using a likelihood test:

$$S(s, I) = \frac{p(s|H_0)}{p(s|H_1)} \begin{cases} \geq \theta \text{ accept } H_0 \\ < \theta \text{ reject } H_0 \end{cases} \tag{1}$$

where θ is a decision threshold determined by the context in which the biometric system is deployed.

We can model the probability $p(s|H_0)$, using the GMM λ_I, as $p(s|\lambda_I)$. The speech signal is converted by the front-end algorithms to a set of K feature vectors, each of dimension D, so we can write:

$$p(s|H_0) = p(\{\boldsymbol{x}_1, .., \boldsymbol{x}_K\} |\lambda_I) = \prod_{j=1}^{K} p(x_j|\lambda_I) \tag{2}$$

The probability that a D-dimensional feature vector $\boldsymbol{x}$ derives from the GMM λ is shown by the following equation:

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{N} w_i p_i(\boldsymbol{x}) \tag{3}$$

where N is the number of components of λ, and w_i and p_i are, respectively, the weight and the individual probability density of the i-th component. Each p_i can be written as:

$$p_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\mu_i)'\Sigma_i(\boldsymbol{x}-\mu_i)} \tag{4}$$

978-1-4244-6304-6/10/$26.00 ©2010 IEEE

where μ_i and Σ_i are respectively the mean vector and the covariance matrix of p_i.

So the λ GMM is defined by the following parameters:

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \tag{5}$$

Those three parameters are estimated using the Expectation-Maximization algorithm [4] in the training phase.

In order to compute the score (1) that must be compared to the system's threshold, we still need to estimate $p(s|H_1)$. This can be done by building a model of all the expected speech signals, a so-called background model (λ_W) [5].

So the final score of the verification process, expressed in terms of log-likelihood ratio, is

$$\Lambda(s) = \log S(s, I) = \log p(s|\lambda_I) - \log p(s|\lambda_W) \tag{6}$$

Our implementation of a GMM system is based on the ALIZE/SpkDet toolkit [6]

III. Speech and background noise database

The database used for the tests was extracted from the TIMIT speech corpus. From the 8 different geographical regions represented by the TIMIT, only the DR1 subset - composed by speakers coming from New England – was chosen, and from it 38 people (half male and half female) were selected.

The 10 sentences spoken by each person, sampled at 8 kHz and linearly quantized using 16 bits per sample, have been used to produce a clean conversation composed by talkspurt segments (ON) normalized to an average power level of $-26dB_{ovl}$ and silence segments (OFF). The ON-OFF statistics were chosen using the model proposed in [7].

The noise database comprises a set of recordings of different types of background noise each lasting 3 minutes, sampled at 8 kHz and linearly quantized using 16 bits per sample. The types of noise contained in the database fall into the following categories:

- **Car**, recordings made inside a car;
- **Office**, recordings made inside an office during working hours;
- **Factory**, recordings made inside a factory;
- **Construction**, recordings of the noise produced by the equipment used in a building site;
- **Train**, recordings made inside a train;

For each type of noise, the clean sequence was digitally summed to the noise, in order to get sequences with four different real SNRs in the activity segments: 0, 10 and 20 dB.

IV. Performance Evaluation and Results

In order to verify the performance of our approach, we computed the genuine match scores and the impostor match scores for different types of noises and signa-to-noise ratio (SNR). In particular, we computed the False Match Rate (FMR) and the False Non-Match Rate (FNMR) against different decision thresholds values. The plot of FNMR against FMR, called Detection Error Tradeoff (DET), is shown in the following figures.

Figg. 1 compare the performance on the basis of noise type for: (a) SNR=20 dB, (b) SNR=10 dB, (c) SNR=0 dB. In all cases we can notice major performance degradation after raising the noise level volume and a different impact on the system performance made by various noise types. In particular, car noise has less impact (EER=13 %) while construction noise is the most degradating noise type (EER=24 %). Algorithm performance in clean sequences points out EER value about 8 %, so the impact of the noise compromises the performance for EER percentage basis ranging from 5 to 15 %.

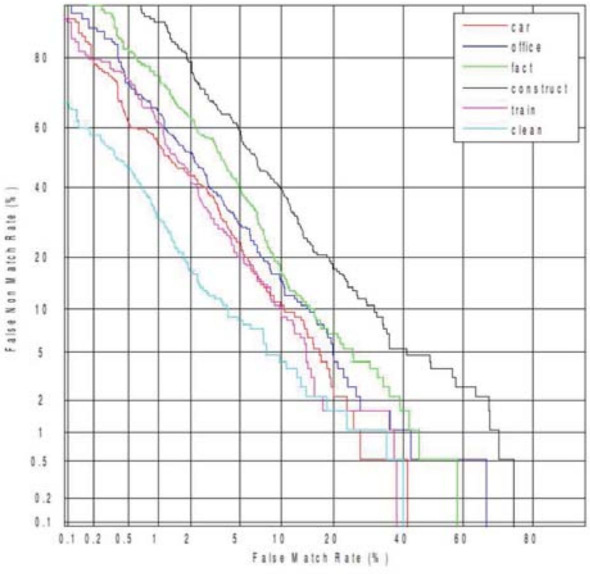
Another important result regards the performance on the basis of the duration of the sequences both during model training phase and that of the test. Fig. 2 compares DET achieved between clean sequences in two cases of the application of:

- 2 training sequences (duration 6,24 sec), 2 true test sequences (duration 6,24 sec) and 2 false test sequences (6,24 sec);
- 8 training sequences (duration 25 sec), 2 true test sequences (6,24 sec) and 2 false test sequences (6,24 sec);

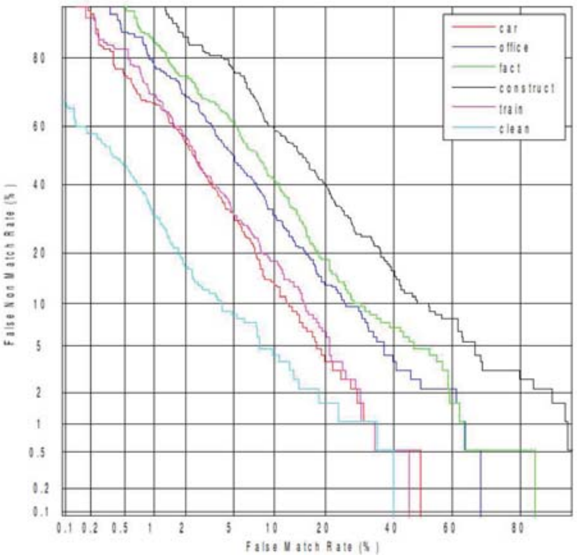
In this case the real impact of the training duration on the total system performance is underlined.

Fig. 3 shows the opposite case where a different duration of the test sequences is applied, in particular:

- 2 training sequences (duration 6,24 sec), 3 true test sequences (duration 9,36 sec) and 3 false test sequences (9,36sec);



(a)



(b)

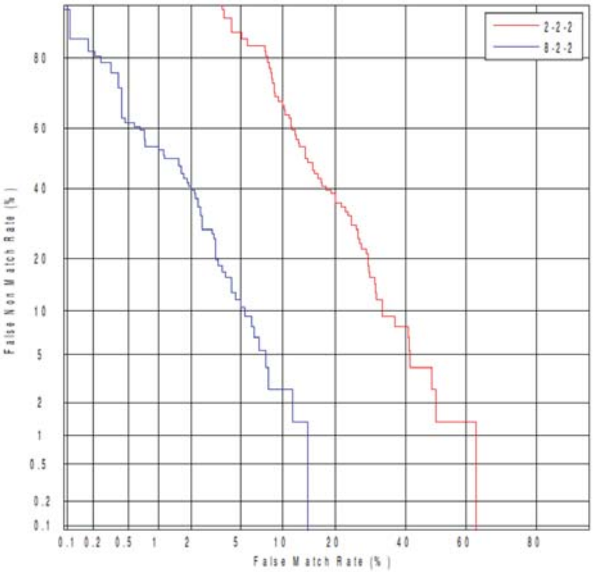
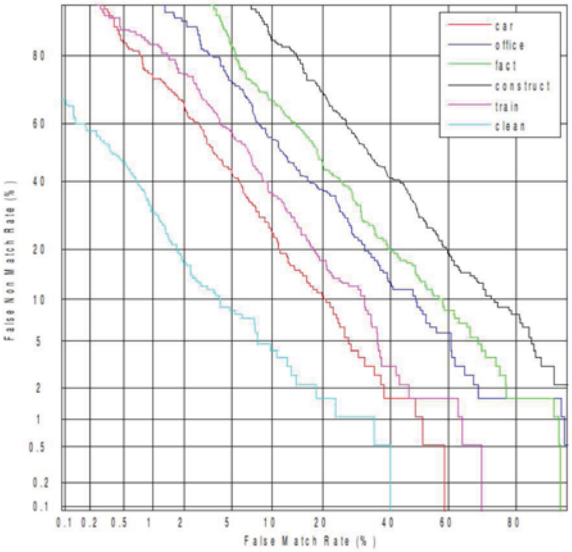


Figure 2



(c)

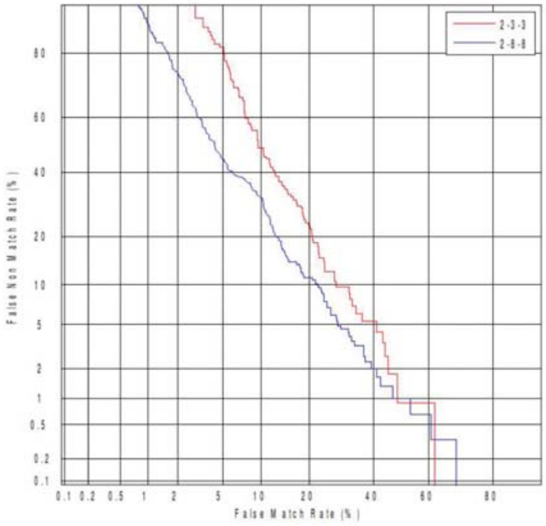


Figure 3

- 2 training sequences (6,24 sec), 8 true test sequences (25 sec) and 8 false test sequences (25 sec).

In this case the various duration of the test sequences does not have much impact and the performances are very similar. Therefore, from this result emerges that, for automatic speaker recognition, it is better to apply longer duration sequences for training and shorter duration sequences for testing.

Finally, Fig. 4 compares system performance in three different phases: comparison of clean type training and testing sequences, comparison of clean training sequence and degraded testing sequence by car noise with SNR 0dB, and comparison of training and testing sequences both degraded by car noise with SNR 0dB. Analysing three DET curves it is possible to see that the application of one noisy sequence also for training does not contribute to the improvement of the performance which remains similar as in clean-noisy case. Generally, speech biometric system performance depends on the degradation of one of the compared sequences (phonic test and testing).

V. Conclusions

The paper presents a research on the robustness, on duration and natural noises, of the automatic systems used for the identification of a person in a forensic setting. The results show that the noise can have the impact on the degradation of the performance (see EER values) which vary from 100 % to 300 % on the basis of the type of noise which depends on only one of two compared sequences.

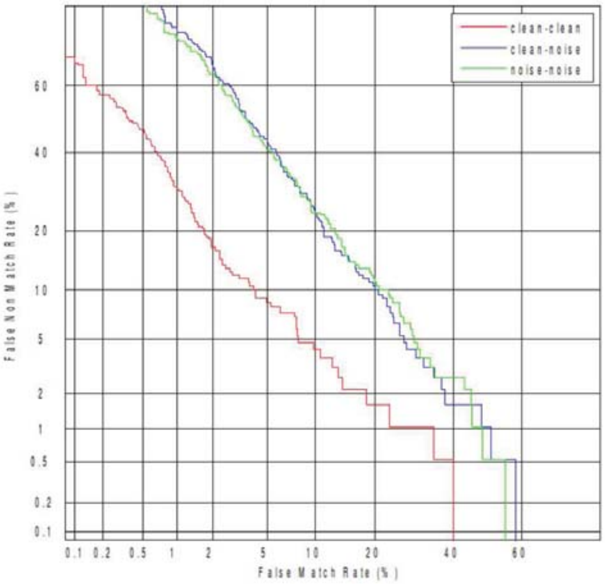


Figure 4

The duration of the sequences is a very important parameter, mostly for training phase, for which it is necessary to have at least 25 seconds long talk.

References

[1] J. Richiardi and A. Drygajlo, "Evaluation of speech quality measures for the purpose of speaker verification," in *Proceedings of Odyssey, The Speaker and Language Recognition Workshop*, 2008.

[2] M. Falcone, A. Paoloni, and N. De Sario, "Idem: A software tool to study vowel formant in speaker identification," in *Proceedings of the ICPhS*, 1995, pp. 145–150.

[3] F. Beritelli, "Effect of background noise on the snr estimation of biometric parameters in forensic speaker recognition," in *Proceeding of the International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2008.

[4] Geoffrey J. McLachlan and Thriyambakam Krishnan, *The EM Algorithm and Extensions*, Wiley, 1997.

[5] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, 2000, p. 2000.

[6] Jean-François Bonastre, Nicolas Scheffer, Driss Matrouf, Corinne Fredouille, Anthony Larcher, Re Preti, Gilles Pouchoulin, Nicholas Evans, Benoît Fauve, and John Mason, "Alize/Spkdet: a state-of-the-art open source software for speaker recognition," .

[7] P. T. Brady, "A model for generating on-off speech patterns in two-way conversation," *Bell Syst. Tech. J.*, pp. 2445–2472, September 1969.

[8] Douglas A. Reynolds and Richard C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, 1995.