Chapter 6

Performance Evaluation of Automatic Speaker Recognition Techniques for Forensic Applications

Francesco Beritelli and Andrea Spadaccini

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/52000

1. Introduction

Speaker recognition is a biometric technique employed in many different contexts, with various degrees of success. One of the most controversial usage of automatic speaker recognition systems is their employment in the forensic context [1, 2], in which the goal is to analyze speech data coming from wiretappings or ambient recordings retrieved during criminal investigation, with the purpose of recognizing if a given sentence had been uttered by a given person.

Performance is one of the fundamental aspects of an FASR (Forensic Automatic Speaker Recognition) system. It depends strongly on the variability in the speech signal, noise and distortions in the communications channel. The recognition task faces multiple problems: unconstrained input speech, uncooperative speakers, and uncontrolled environmental parameters. The speech samples will most likely contain noise, may be very short, and may not contain enough relevant speech material for comparative purposes. 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 recognition phase [3, 4]. Each of these variables makes reliable discrimination of speakers a complicated and daunting task.

Typically the performance of a biometric system is determined by the errors generated by the recognition. There are two types of errors that can occur during a verification task: (a) false acceptance when the system accepts an imposter speaker; and (b) false rejection when the system rejects a valid speaker. Both types of errors are a function of the decision threshold. Choosing a high threshold of acceptance will result in a secure system that will accept only a few trusted speakers, however, at the expense of high false rejection rate (FRR) or False Non Match Rate (FNMR). Similarly choosing a low threshold would make the system more user friendly by reducing false rejection rate but at the expense of high false acceptance rate (FAR) or False Match Rate (FMR). This trade-off is typically depicted using a decision-error trade-off (DET) curve. The FAR and FRR of a verification system define different operating points on the DET curve.

In general, to understand what are the causes that contribute most to the total error of an FASR system, it is important to evaluate the performance of individual blocks or phases of the system. Knowing the impact on the performance of individual subsystems of a speaker recognition algorithm (manual, semiautomatic or automatic) allows us to understand what aspects should be better cared for if you want to achieve the performance targets required by the FASR system.

As of the writing of this document, semi-automatic speaker recognition techniques are still employed in Italian courts; this means that an expert witness analyzes the speech data with the aid of some *ad hoc* software, that usually gives the freedom to change some parameters that can affect the final outcome of the identification.

It is obvious that human errors can lead to wrong results, with disastrous consequence on the trial.

In this chapter, we will analyze how efficiently and reliably can state-of-the-art speaker recognition techniques be employed in this context, what are their limitations and their strengths and what must be improved in order to migrate from old-school manual or semi-automatic techniques to new, reliable and objective automatic methods.

It is well-known that speech signal quality is of fundamental importance for accurate speaker identification [5]. The reliability of a speech biometry system is known to depend on the amount of available data, in particular on the number of vowels present in the sequence being analysed, and the quality of the signal [6]; the former affects the resolution power of the system, while the latter impacts the correct estimation of biometric indexes.

In this chapter, we will analyze the behaviour of some speaker recognition techniques when the environment is not controlled and the speech sequences are disturbed by background noise, a very frequent condition in real-world forensic data.

This chapter is organized as follows: in Section 2 we will describe the baseline speaker recognition systen and the speech and noise databases used for the experiments; in Section 3 we will analyze the performance of two Signal-to-Noise (SNR) estimation algorithms; in Section 4 we will analyze the performance of a speaker recognition toolkit; in Section 5 we will analyze the impact of Voice Activity Detection (VAD) algorithms on the recognition rates; finally, in Section 6 we will draw our conclusions.

2. Baseline speaker recognition system and speech/noise databases

2.1. Alize/LIA_RAL

The speaker verification system used in all the experiments is based on ALIZE/LIA_RAL , that is described in this section.

The ALIZE/LIA_RAL toolkit is developed jointly by the members of the ELISA consortium [8], and consists of two separate components: Alize, that is the low-level statistical framework, and LIA_RAL, that is the set of high-level utilities that perform each of the tasks of a state-of-the-art speaker recognition system. The latter is also sometimes referred to as Mistral[9].

One of its main advantages is the high level of modularity of the tools: each program does not directly depend on the others and the data between the modules is exchanged via text

http://dx.doi.org/10.5772/52000

files whose format is simple and intuitive. This means that researchers can easily change one of the components of the system with their own program, without having to modify its source code but only adhering to a set of simple file-based interfaces.

In this section, we will briefly describe all the components of a typical experiment that uses the ALIZE/LIA_RAL toolkit.

2.1.1. Feature extraction

LIA_RAL does not contain any module that performs feature extraction; all the experiments in this chapter used the Speech Signal Processing Toolkit (SPro) [10] for feature extraction tasks. SPro allows to extract different types of features, using filter-banks, cepstral analysis and linear prediction.

2.1.2. Frames selection

The second step of the recognition process is to remove the frames that do not carry useful information. When dealing with the speech signal, this task is carried on by VAD algorithms, some of which will be described in Section 5. Each of these VAD algorithms was implemented by a different program, and their output was always converted to a format that is compatible with the LIA_RAL toolkit.

The default VAD algorithm in LIA_RAL, described in Section 5.3, is implemented in the utility *EnergyDetector*.

2.1.3. Feature normalization

The third step is the feature normalization, that changes the parameters vectors so that they fit a zero mean and unit variance distribution. The distribution is computed for each file.

The tool that performs this task is called *NormFeat*.

2.1.4. Models training

To use the UBM/GMM method [11], it is necessary to first create a world model (UBM), that represents all the possible alternatives in the space of the identities enrolled in the system; then, from the UBM, the individual identity templates are derived from the UBM using the Maximum A-Posteriori (MAP) estimation algorithm.

The tool used for the computation of the UBM is *TrainWorld*, while the individual training models are computed using *TrainTarget*.

2.1.5. Scoring

The computation of scores is done via the *ComputeTest* program, that scores each feature set against the claimed identity model and the UBM, and gives as output the log-likelihood ratio.

In order to take a decision, the system has then to compare the score with a threshold, and then accept or reject the identity claim. The decision step is implemented in LIA_RAL utility *Scoring*, but in this chapter we have not used it.

2.2. The TIMIT speech database

All the speaker recognition experiments described in this chapter use as a speech database a subset of the TIMIT (Texas Instrument Massachusetts Institute of Technology) database, that will be briefly described in this section.

The TIMIT database contains speech data acquired from 630 people, that are split in subsets according to the Dialect Region to which each of them belongs. Each DR is further split in training and test set. The number of speakers contained in each DR, and their division in training and test set are reported in Table 1.

Dialect Region	Total	Training	Test
New England (DR1)	49	38	11
Northern (DR2)	102	76	26
North Midland (DR3)	102	76	26
South Midland (DR4)	100	68	32
Southern (DR5)	98	70	28
New York City (DR6)	46	35	11
Western (DR7)	100	77	23
Moved around (DR8)	33	22	11
Totals:	630	462	168

Table 1. Composition of the TIMIT data set

This database was explicitly designed to provide speech researchers with a phonetically rich dataset to use for research in speech recognition, but it is widely adopted also in the speaker recognition research community.

It contains three types of sentences, dialectal (SA), phonetically-compact (SX) and phonetically diverse (SI). The total number of spoken sentences is 6300, 10 for each of the 630 speakers. There is some superposition between speakers, because there are sentences that are spoken by more than one person. Each person, however, has to read 2 SA sentences, 5 SX sentences and 3 SI sentences.

The database also contains annotations about the start and end points of different lexical tokens (phonemes, words and sentences). This was especially useful for the research on SNR and VAD, because we could compare our algorithms with ground truth provided by the database itself.

2.3. The noise database

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;

http://dx.doi.org/10.5772/52000

- **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;

3. Performance evaluation of SNR estimation algorithms

In forensics, one of the most widely adopted methods to assess the quality of the intercepted signal is based on the estimation of the Signal to Noise Ratio (SNR), that should not be lower than a critical threshold, usually chosen between 6 and 10 dB [6]. It is possible to estimate the SNR using manual or semi-automatic methods. Both of them exploit the typical ON-OFF structure of conversations, which means that on average there are times when there is a speech activity (talkspurt) and times when nobody is talking, and the signal is mainly composed by environmental noise recorded by the microphone (background noise). With manual methods, the SNR is estimated choosing manually the segment of talkspurt and the segment of background noise immediately before or after the talkspurt. The estimation is computed by the following formula:

$$SNR_{est} = \frac{P_{talk} - P_{noise}}{P_{noise}} \tag{1}$$

Semi-automatic estimation methods use a Voice Activity Detection (VAD) algorithm that separates the ON segments from the OFF segments in a given conversation, and use those segments to estimate the SNR [12, 13].

Both algorithms do not give an exact value of the SNR, because the noise sampled for the SNR estimation is different from the noise that degraded the vocal segment for which the SNR is being estimated. This happens because the noise level can be measured only when the speakers are not talking, in an OFF segment.

Sometimes the estimation error causes the elimination of good-quality data (under-estimation of the SNR), while sometimes it causes the usage of low-quality biometric data that was probably corrupted by noise in the subsequent identification process (over-estimation of the SNR)

In this section, we will discuss about the accuracy of the SNR estimation methods, comparing their average estimation error to the real SNR.

3.1. Speech and background noise database

In this experiment, we used speech data coming from 100 people, half female and half male, randomly selected from the DR1 subset of the TIMIT database

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 [14].

We used 4 kinds of background noise: Car, Office, Stadium, Construction.

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, 20 and 30 dB.

3.2. SNR estimation methods

Both analyzed SNR estimation methods exploit the manual phonetic marking offered by the TIMIT database. In particular, for the sake of simplicity, we selected a restricted subset of vowel sounds ("ae", "iy", "eh", "ao"), of which only the central 20 ms were considered.

The manual SNR estimation method computes the estimated SNR as the ratio between the power of the signal of the current vowel, (P_{talk}) lasting 20ms, to the power of noise, (P_{noise}), measured at the nearest OFF segment and lasting 20 ms. The classification of the signal in ON and OFF segments is done manually.

The semi-automatic method uses the VAD algorithm to automatically classify ON and OFF segments. The VAD used is the one standardized by the ETSI for the speech codec AMR [15]. In this case, the noise power is measured using the nearest OFF segment classified by the VAD.

The values obtained by the two algorithms have then been compared to the real SNR, computed as the ratio between the power of the vowel measured on the clean signal and the power of the background noise measured in the same temporal position but on the noise sequence.

3.3. Results

The first analysis that we present is the computation of the average estimation errors. In each subplot, two axis represent the SNR and the vowel, while the third one represents the average estimation error.

Figure 1 shows the average estimation error for the manual method, while Figure 2 shows the same error, but for the semi-automatic method.

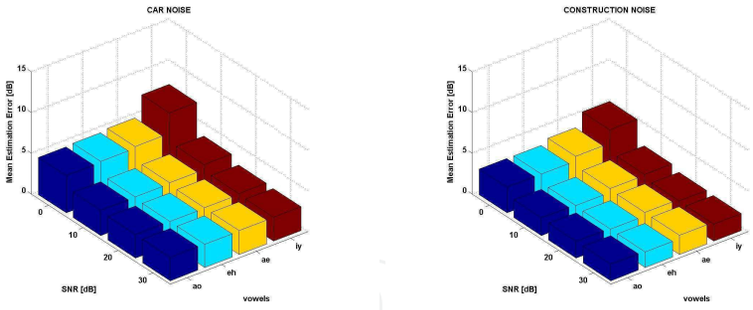
The performance of both methods are similar for the Car and Noise, with an average error between 3 and 5 dB of difference with the reference SNR.

A comparison of the errors reveals that the usage of the automatic method increases the average error by 1 dB in case of the Car, Construction and Office noises, while the increase is larger (between 2 and 5 dB) for the Stadium noise.

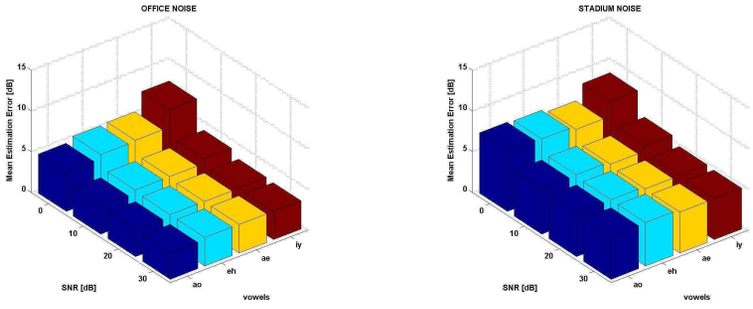
Even though the VAD impact on the SNR estimation depends on the type of noise, it however does not lead to heavily poorer performance because on average the error grows by only 1-2 dB.

In both cases, when the reference SNR is 0 dB it can be seen that the "iy" vowel is subject to a high sensitivity for each kind of noise. The average estimation error generally is larger by 20-30% with respect to the other vowels.

http://dx.doi.org/10.5772/52000

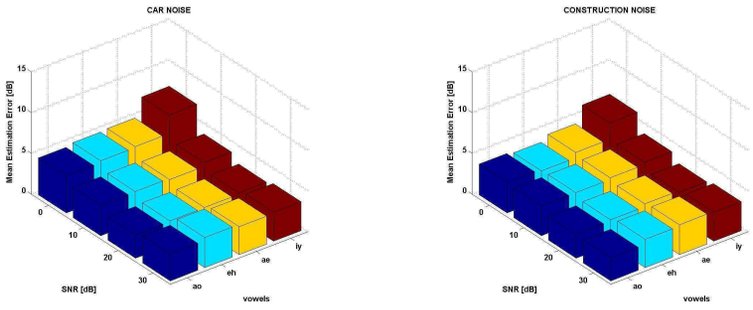


(a) Car, Construction

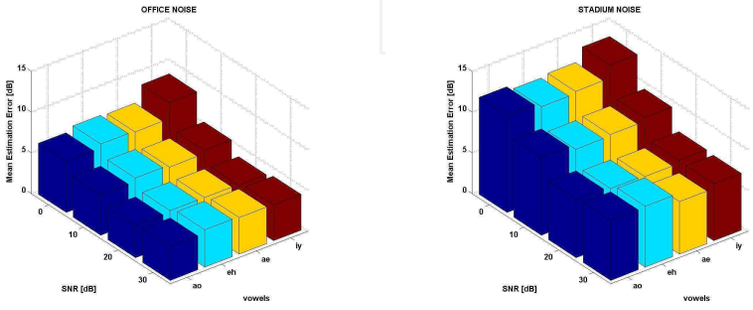


(b) Office, Stadium

Figure 1. Average SNR estimation errors for the manual method

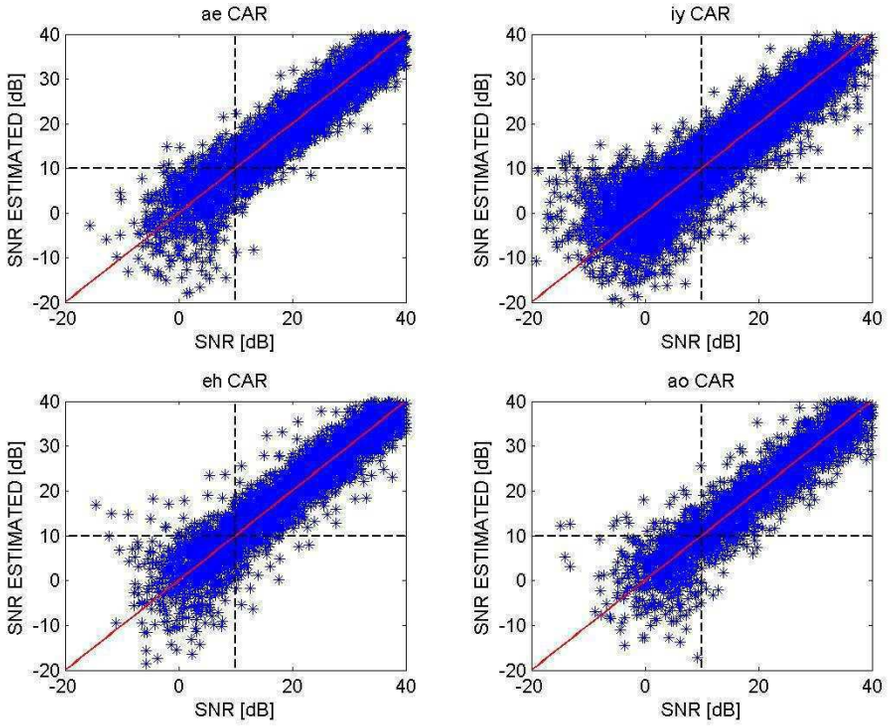


(a) Car, Construction

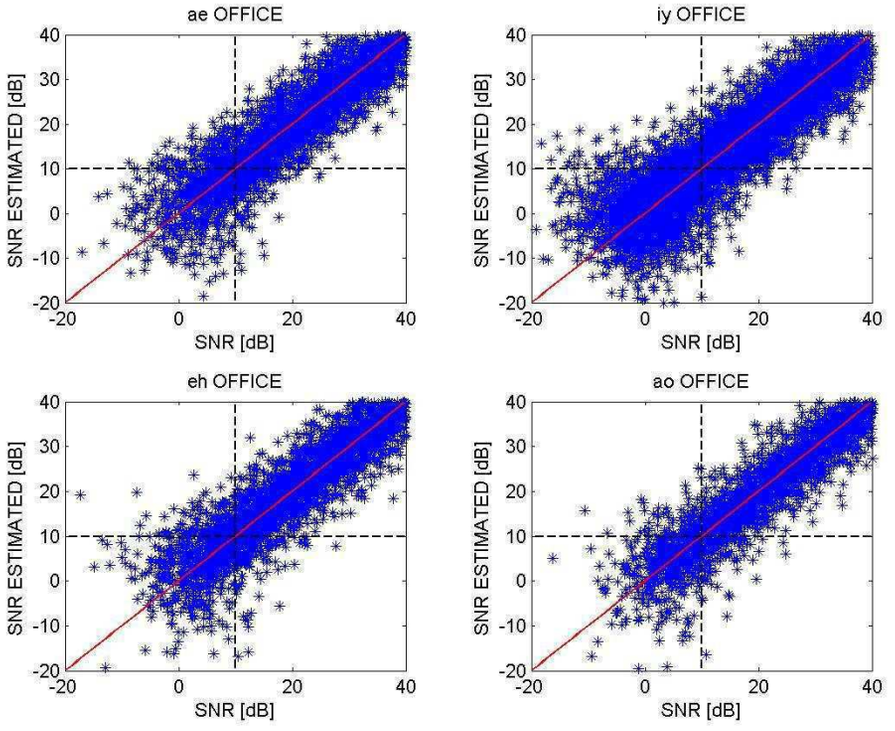


(b) Office, Stadium

Figure 2. Average SNR estimation errors for the semi-automatic method



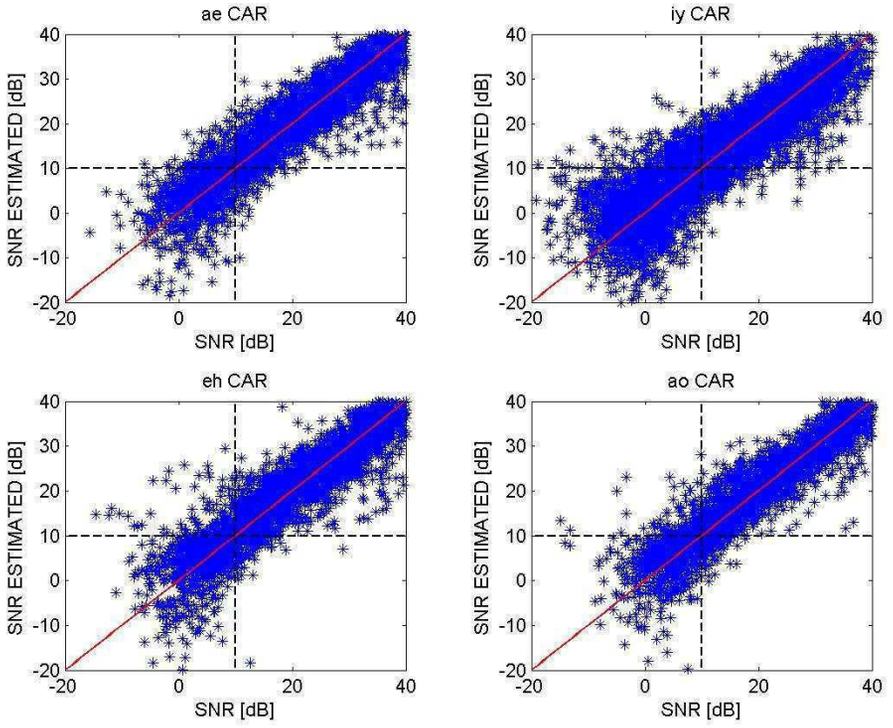
(a) Car



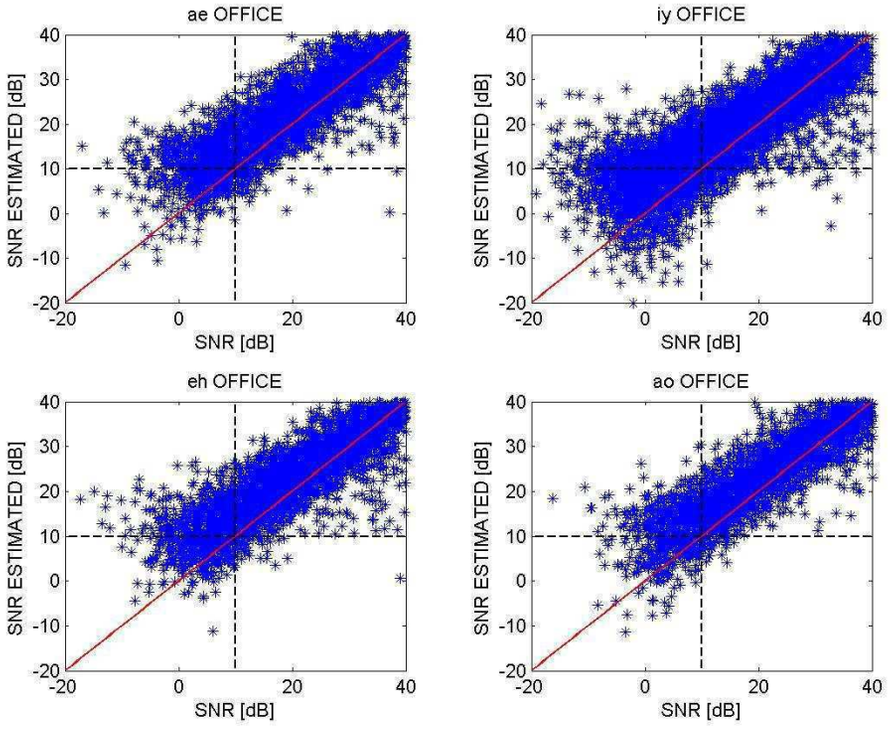
(b) Office

Figure 3. Real vs. estimated SNR, manual method

http://dx.doi.org/10.5772/52000



(a) Car



(b) Office

Figure 4. Real vs. estimated SNR, semi-automatic method

The plots in Figure 3 and Figure 4 show the correlation between the real SNR and the estimated SNR for each of the 4 vowels in case of Car and Office noise. If we assume a critical threshold for rejecting a biometric sample of 10 dB, it is possible to outline 4 regions in each of these plots: the upper-left one, that encompasses data erroneously used because the SNR was over-estimated; the lower-right region, that comprises data erroneously discarded because the SNR was under-estimated, and the remaining regions (upper-right and lower-left), that contain data that were correctly discarded or used for the subsequent identity verification phases.

Car noise	**ae**	**iy**	**eh**	**ao**
Bad data used	15.39%	11.63%	15.34%	16.82%
Good data discarded	5.49%	6.75%	4.49%	4.91%
Office noise	**ae**	**iy**	**eh**	**ao**
Bad data used	22.14%	14.95%	21.76%	17.70%
Good data discarded	5.97%	7.97%	6.41%	6.00%

Table 2. Percentage errors for the manual method

Car noise	**ae**	**iy**	**eh**	**ao**
Bad data used	18.56%	15.42%	18.11%	18.77%
Good data discarded	4.94%	6.86%	4.61%	4.33%
Office noise	**ae**	**iy**	**eh**	**ao**
Bad data used	60.45%	42.70%	58.55%	56.46%
Good data discarded	2.35%	3.28%	1.53%	1.59%

Table 3. Percentage errors for the semi-automatic method

Tables 2 and 3, respectively, for manual and semi-automatic methods, show the error percentages depicted in Figure 3 and Figure 4. The semi-automatic method induces an increment of the percentage of low-quality data that is used for subsequent elaboration for the Office noise, while the percentages for the Car noise are similar to the ones of the manual method.

In the end, comparing the percentage of low-quality data erroneously used, it can be deduced that each vowel reacts in different ways: for instance, the “iy” vowel is one of the most robust. A similar comparison can be carried out in terms of high-quality data erroneously discarded.

4. Performance evaluation of Alize-LIA_RAL

In this section we present a study on how a speaker recognition system based on the Alize/LIA_RAL toolkit behaves when the data is affected by background noise. In particular, the section shows both the performance using a "clean" database and the robustness to the degradation of various natural noises, and their impact on the system. Finally, the impact of the duration of both training and test sequences is studied.

http://dx.doi.org/10.5772/52000

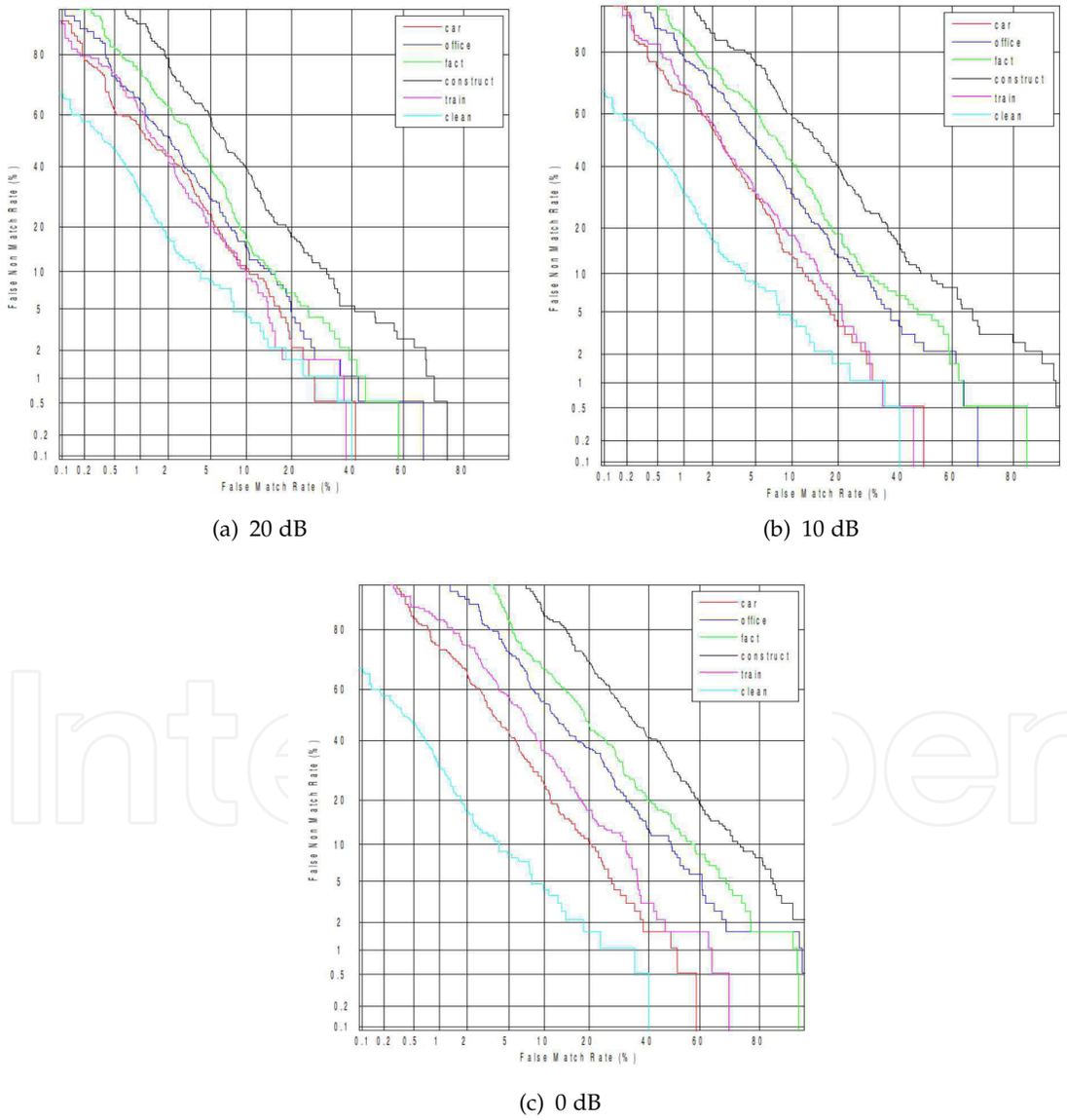
4.1. Speech and background noise database

For this experiment, we used the training portion of the DR1 TIMIT subset, that contains 38 people.

We generated the clean and noisy databases using the same protocol described in Section 3.1.

4.2. Performance evaluation and results

In order to verify the performance of our system, we computed the genuine match scores and the impostor match scores for different types of noises and signal-to-noise ratio (SNR). The Detection Error Trade-off of each test case is shown in the following figures.



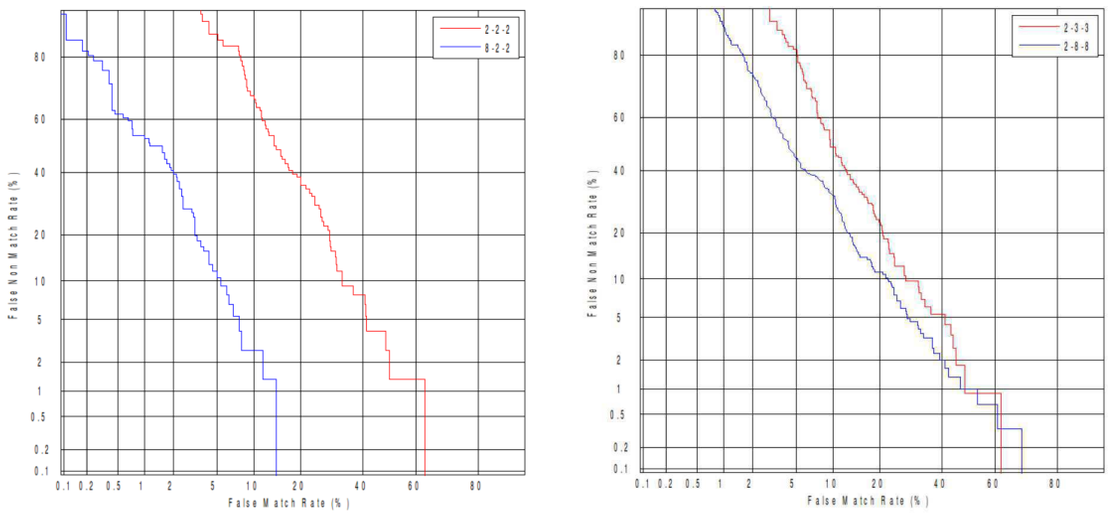
(a) 20 dB

(b) 10 dB

(c) 0 dB

Figure 5. DET vs. Noise type

Figures 5 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 degrading noise type (EER=24 %). Algorithm performance in clean sequences points out an EER value of about 8 %, so the impact of the noise compromises the performance for EER percentage basis ranging from 5 to 15 %.



(a) 2-2-2 vs. 8-2-2 (b) 2-3-3 vs. 2-8-8

Figure 6. Training length vs. test length

Another important result is the discovered relation about the recognition performance and the duration of the training and testing sequences. Figure 6(a) compares the DET achieved using clean sequences spanning the following durations:

- 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 evident.

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

http://dx.doi.org/10.5772/52000

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

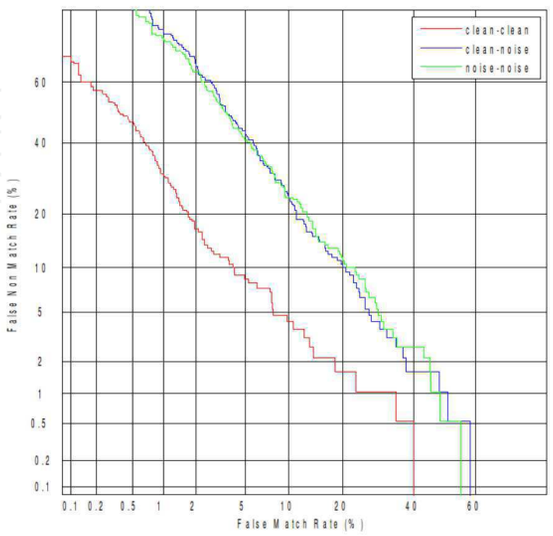


Figure 7. Clean-clean, Clean-Noisy, Noisy-Noisy

Finally, Figure 7 compares system performance in three different modalities: 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 the three DET curves it is possible to see that employing one noisy sequence in the training phase does not contribute to the improvement of the performance, which remains similar to the clean-noisy case. Generally, we can therefore conclude that speaker identification performance is sensitive to the degradation of one of the compared sequences (phonic test and testing).

5. Performance evaluation of voice activity detection algorithms

The performance of biometric speaker verification systems is largely dependent on the quality level of the input signal [11]. One of the most important components of such a system is the Voice Activity Detection (VAD) algorithm, as it has the duty of separating speech frames and noise frames, discarding the latter and feeding the speech frames to the rest of the system. This task becomes quite challenging as the Signal-to-Noise Ratio (SNR) of the input signal goes down [12] [13].

A VAD algorithm can use many techniques to classify speech and noise, such as an energy threshold or the analysis of the spectral characteristics of the audio signal. Due to these differences, different algorithms can behave differently in a given noise condition, and this is the reason for the study presented by this section.

The context in which we operate is the analysis of phone tappings in forensic investigations, and our task is to determine whether the conversation was carried on by a suspect or not. Those tappings are often noisy, so we generated from a speech database some audio files

with the typical ON-OFF statistics of phone conversations and artificially added to them background noise in order to evaluate the performance of VAD algorithms and speaker identification at different SNR levels[3].

Our objective is to demonstrate that the usage of a single VAD is not the optimal solution, however, biometric identification performance can be improved by introducing a noise estimation component that can dynamically choose the best VAD algorithm for the estimated noise condition.

5.1. The speech database

For our task we selected a TIMIT subset composed by 253 speakers, namely the union of DR sets 1, 2 and 3. Of those speakers, 63 were destined to train the UBM and 190 were used to train the identity models and to compute the match scores. With those speakers, we obtained 190 genuine match scores and 35910 ($190 \cdot 189$) impostor match scores for each simulation.

The speech files were used to generate two one-way conversation audio files, each containing speech material from 5 speech files and with an activity factor of 0.4, using the algorithm described in Section 5.2. In the case of the UBM speakers, both sequences were processed for the training phase, while in the case of identity models one sequence was used for the model training and the other was used for the computation of match scores.

The whole database was downsampled to 8kHz, to better match the forensic scenario, and normalized to an average power level of -26dB_{ovl}.

5.2. Generation of one-way conversations

In order to simulate the forensic scenario, and to give realistic input data to the VAD algorithms, we generated for each speaker two audio files that mimic one side of a two-people conversation, inserting speech and pauses according to the model described in [16], that will now be briefly described.

According to this model, a conversation can be modelled as a Markov chain, whose state can be one of the following: A is talking, B is talking, Mutual silence, Double talk. A and B are the two simulated speakers.

The chain is depicted in Figure 8, along with the transition probabilities between the states.

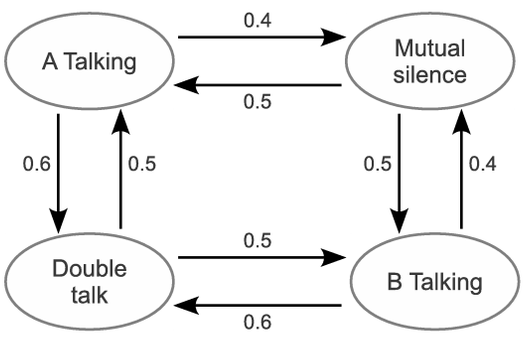


Figure 8. Markov chain used to generate the conversations

http://dx.doi.org/10.5772/52000

The permanence in each of these states is given by the following equations:

$$T_{st} = 0.854 \ \ln(1 - x_1)$$
$$T_{dt} = 0.226 \ \ln(1 - x_2)$$
$$T_{ms} = 0.456 \ \ln(1 - x_3)$$

where $0 < x_1, x_2, x_3 < 1$ are random variables with uniform distribution. T_{st} is the permanence time in the states in which a single speaker is talking, T_{dt} is associated to the double talk state and T_{ms} is used for the mutual silence state.

This model represents a two-way conversation, but we are interested in generating speech for one of the two sides of the conversation. So when the model is in the state "A is speaking" or "Mutual talk", the generator adds speech material to the output sequence, while in the other two states the generator adds silence.

For this experiment, we used the Car, Office and Factory noises.

5.3. The LIA_RAL VAD

The LIA_RAL VAD is an energy-based off-line algorithm that works by training a GMM on the energy component of the input features. It then finds the Gaussian distribution with the highest weight w_i and uses its parameters to compute an energy threshold according to the following formula:

$$\tau = \mu_i - \alpha\sigma_i$$

where α is a user-defined parameter, and μ_i and σ_i are the parameters of the selected gaussian mixture Λ_i.

The energy threshold is then used to discard the frames with lower energy, keeping only the ones with a higher energy value.

5.4. The AMR VAD

The Adaptive Multi-Rate (AMR) Option 1 VAD [15] is a feature-based on-line algorithm that works by computing the SNR ratio in nine frequency bands, and decides which frames must be kept by comparing the SNRs to band-specific thresholds.

Note that this VAD is not optimized for speaker verification tasks, as it has the objective of minimizing the decision time, and it is designed to be used in real-time speech coding applications, while in a forensic biometric system the delay is not a significant parameter to minimize, and thus the VAD could use information from all the input signal to make its decision, as the LIA_RAL VAD does.

5.5. Evaluating VAD performance

In order to evaluate the VAD performance, we need to compare the results of the classification on a given input signal with a reference ideal classification that we know for sure to be correct.

In our experimental set-up, this ideal classification is derived by labelling the start and the end of speech segments generated by the model described in Section 5.2. This classification does not take into account pauses that can occur during the TIMIT spoken sentences, but it is a good approximation of an ideal classification.

The VAD classifier can misinterpret a given input frame in two ways: detecting a noise frame as speech (Noise Detected as Speech, NDS) or classifying a speech frame as noise (Speech Detected as Noise, SDN).

Those two errors are then further classified according to the position of the error with respect to the nearest word; see [17] for a discussion of those parameters.

For our analysis, we are not interested in the time when the misclassification occurs, as it is mainly useful when evaluating the perception effects of VAD errors [18], so we use the two NDS and SDN parameters, defined as follows for a single conversation:

$$NDS_{\%} = \frac{N_{NDS} \cdot f}{C}$$
$$SDN_{\%} = \frac{N_{SDN} \cdot f}{C}$$

where N_{NDS} and N_{SDN} are, respectively, the number of NDS and SDN frames, f is the frame length expressed in seconds and C is the duration of the conversation expressed in seconds.

We then define a Total Error Rate (TER), as:

$$TER_{\%} = NDS_{\%} + SDN_{\%} \tag{2}$$

The TER is the percentage of audio frames that are misclassified by the VAD.

5.6. Experimental results

The starting point of our experiments is the creation of 9 noisy speech databases, obtained by summing to the one-way conversation speech database described in Section 5.2 the Car, Office and Factory noises, artificially setting the SNR to 20 dB, 10 dB and 0 dB.

Next the Equal Error Rate (EER) was computed over each database, first with the ideal segmentation and then by swapping this segmentation with the ones generated by the LIA_RAL VAD and by the AMR VAD, for a total of 27 simulations.

Finally, the VAD errors were computed using the metrics defined in Section 5.5.

http://dx.doi.org/10.5772/52000

In the clean case (Table 4), we reported the average $NDS_{\%}$, $SDN_{\%}$ and $TER_{\%}$, computed over all the speech samples used to run the speech verification simulations, one time for each VAD algorithm (including the ideal VAD).

In the noisy cases (Tables 5, 6, 7), since for each VAD the simulation was run once for each SNR level, the reported VAD error metrics are the average of the average of the value of each metric (denoted with μ) computed over all the speech samples, and their standard deviations σ are reported in order to better understand the nature of the data presented. Obviously, the VAD errors of the ideal VAD are always zero, so the standard deviation is omitted from the tables.

5.7. Analysis of the results

VAD algorithm	**EER (%)**	$\overline{NDS_{\%}}$	$\overline{SDN_{\%}}$	$\overline{TER_{\%}}$
ideal	**3.76**	0	0	0
AMR	4.36	1.08	4.81	5.89
LIA_RAL	3.77	4.33	31.74	36.07

Table 4. Results for clean speech

	EER (%)			**VAD Errors ($\mu \pm \sigma$, %)**		
VAD	0dB	10dB	20dB	$\overline{NDS_{\%}}$	$\overline{SDN_{\%}}$	$\overline{TER_{\%}}$
ideal	5.77	3.60	3.55	0	0	0
AMR	**4.95**	**4.77**	**3.38**	7.30 ± 1.55	4.88 ± 0.34	12.18 ± 1.89
LIA_RAL	6.49	5.43	4.87	3.95 ± 0.12	34.38 ± 0.24	38.33 ± 0.36

Table 5. Results table for CAR noise

	EER (%)			**VAD Errors ($\mu \pm \sigma$, %)**		
VAD	0dB	10dB	20dB	$\overline{NDS_{\%}}$	$\overline{SDN_{\%}}$	$\overline{TER_{\%}}$
ideal	8.52	5.69	4.10	0	0	0
AMR	9.77	5.39	**3.75**	41.23 ± 5.24	5.08 ± 0.30	46.31 ± 5.53
LIA_RAL	**6.97**	**5.21**	4.00	0.83 ± 0.85	19.39 ± 3.52	20.22 ± 4.37

Table 6. Results table for OFFICE noise

Looking at Table 4, the first question is why the ideal segmentation yields an EER that is very close to the one that the LIA_RAL VAD obtained, in spite of a greater $\overline{TER_{\%}}$.

This is because the ideal segmentation does not focus only on vocalized sounds, that are known to carry the information that is needed to determine the identity of the speaker, but rather is an indicator of when the generator described in Section 5.2 introduced speech in the audio sequence. This therefore includes some sounds, like fricatives, that should be left out when doing biometric identity comparisons. This also explains the worse performance

	EER (%)			VAD Errors ($\mu \pm \sigma$, %)		
VAD	0dB	10dB	20dB	$\overline{NDS_{\%}}$	$\overline{SDN_{\%}}$	$\overline{TER_{\%}}$
ideal	7.84	5.01	4.70	0	0	0
AMR	7.27	**5.02**	**3.42**	13.64 ± 1.04	6.12 ± 1.33	19.76 ± 2.49
LIA_RAL	**6.58**	5.93	4.37	3.13 ± 1.44	16.87 ± 3.40	20.00 ± 4.84

Table 7. Results table for FACTORY noise

of the ideal VAD in other cases like OFFICE, FACTORY 0 dB, etc. Analyzing the average errors made by the VAD algorithms, it is clear that the AMR VAD usually tends to be more conservative in the decision of rejection of speech, because its $\overline{NDS_{\%}}$ is always greater than LIA_RAL's; on the other hand, LIA_RAL always has a greater $\overline{SDN_{\%}}$ than AMR, and this means that it tends to be more selective in the decision of classifying a frame as noise.

The results for the CAR noise show that the AMR VAD always performs better than the LIA_RAL VAD in terms of EER, and it is supported by a significantly lower TER.

The OFFICE noise results do not show a clear winner between the two algorithms, as for high SNR the AMR VAD performs better, but as the SNR decreases, the LIA_RAL algorithm outperforms the AMR VAD. A similar pattern can be seen in the FACTORY results.

6. Conclusions and future work

In this chapter, we analyzed many of the problems that currently affect forensic speaker recognition. It is clear from the results of the previous sections that there is still no universal approach for speaker recognition in forensic context, and also that this applies to some of the smaller sub-problems.

More specifically, some ideas for future work in the SNR estimation area are:

- develop more effective SNR estimation algorithms, that can guarantee a lower average error and, most importantly, lower variance;
- estimate SNR in the sub-bands of interest of the main biometric indices adopted [3], typically in the fundamental frequency and in the first three formants;
- identify the critical SNR thresholds for each vowel and for each kind of noise by evaluating the impact of the noise on the whole identification process;
- use automatic environmental noise classifiers that allow to choose an SNR estimation model and critical thresholds tailored to the kind of noise [13] [19]

Regarding the selection of VAD algorithms, in the forensic context, where accuracy is truly important and results can be collected off-line, multiple VAD algorithms with different characteristics could be used, and all the identification decisions computed using them could then be fused using a majority rule or other fusion rules. In those critical kinds of analysis, it would be important that most of the decisions agreed between them, or else the disagreement could be an indicator that the choice of the VAD algorithm has a greater weight than desired.

http://dx.doi.org/10.5772/52000

More broadly, based on the results of the research work described in this chapter, it is clear that both the SNR estimation and VAD algorithm selection problems could benefit from an adaptive approach that first estimates the characteristics of background noise and then select the algorithm that performs better in that context [20].

Author details

Francesco Beritelli* and Andrea Spadaccini

* Address all correspondence to: francesco.beritelli@dieei.unict.it

DIEEI Dipartimento di Ingegneria Elettrica Elettronica e Informatica, University of Catania, Italy

7. References

[1] Philip Rose. *Forensic Speaker Recognition*. Taylor and Francis, 2002.

[2] J.P. Campbell, W. Shen, W.M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf. Forensic speaker recognition. *Signal Processing Magazine, IEEE*, 26(2):95 –103, march 2009.

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

[4] Kichul Kim and Moo Young Kim. Robust speaker recognition against background noise in an enhanced multi-condition domain. *Consumer Electronics, IEEE Transactions on*, 56(3):1684 –1688, aug. 2010.

[5] 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.

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

[7] T. May, S. van de Par, and A. Kohlrausch. Noise-robust speaker recognition combining missing data techniques and universal background modeling. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(1):108 –121, jan. 2012.

[8] The ELISA consortium. The elisa consortium, the elisa systems for the nist'99 evaluation in speaker detection and tracking. *Digital Signal Processing*, 10, 2000.

[9] Eric Charton, Anthony Larcher, Christophe Levy., and Jean-Francois Bonastre. Mistral: Open source biometric platform, 2010.

[10] G. Gravier. SPro: speech signal processing toolkit, 2003.

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

[12] F. Beritelli, S. Casale, and S. Serrano. A low-complexity speech-pause detection algorithm for communication in noisy environments. *European Transactions on Telecommunications*, 15:33–38, January/February 2004.

[13] F. Beritelli, S. Casale, and S. Serrano. Adaptive v/uv speech detection based on acoustic noise estimation and classification. *Electronic Letters*, 43:249–251, February 2007.

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

[15] ETSI. Gsm 06.94, digital cellular telecommunication system (phase 2+); voice activity detector (vad) for adaptive multi rate (amr) speech traffic channels; general description. *Tech. Rep. V. 7.0.0*, February 1999.

[16] ITU-T Recommendation P. 59: Artificial conversational speech, March 1993.

[17] F. Beritelli, S. Casale, and A. Cavallaro. A robust voice activity detector for wireless communications using soft computing. *IEEE J. Select. Areas Commun*, 16:1818–1829, December 1998.

[18] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano. Performance evaluation and comparison of g.729/amr/fuzzy voice activity detectors. *IEEE Signal Processing Letters*, 9:85–88, March 2002.

[19] L. Couvreur and M. Laniray. Automatic noise recognition in urban environments based on artificial neural networks and hidden markov models. In *Proceedings of INTERNOISE*, 2004.

[20] F. Beritelli, S. Casale, A. Russo, and S. Serrano. A speech recognition system based on dynamic characterization of background noise. In *Proceedings of the 2006 IEEE International Symposium on Signal Processing and Information Technology*, pages 914–919, 2006.