ELSEVIER

Speech Communication 40 (2003) 261–276

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Towards improving speech detection robustness for speech recognition in adverse conditions

Lamia Karray [a,*], Arnaud Martin [b]

[a] France Télécom – FTR&D/DIH/IPS, 2, Av. P. Marzin, 22307 Lannion Cedex, France
[b] Université de Bretagne Sud, IUT de Vannes, 8, rue Montaigne, 56000 Vannes, France

Received 25 February 2000; received in revised form 18 February 2002; accepted 19 February 2002

Abstract

Recognition performance decreases when recognition systems are used over the telephone network, especially wireless network and noisy environments. It appears that non-efficient speech/non-speech detection (SND) is an important source of this degradation. Therefore, speech detection robustness to noise is a challenging problem to be examined, in order to improve recognition performance for the very noisy communications. Several studies were conducted aiming to improve the robustness of SND used for speech recognition in adverse conditions. The present paper proposes some solutions aiming to improve SND in wireless environment. Speech enhancement prior detection is considered. Then, two versions of SND algorithm, based on statistical criteria, are proposed and compared. Finally, a post-detection technique is introduced in order to reject the wrongly detected noise segments.

Zusammenfassung

Die Spracherkennungsleistung vermindert sich stark, wenn Spracherkennungssysteme in Telefonnetzen schlechter Übertragungsqualität eingesetzt werden und/oder der Anruf in einer Umgebung störender Nebengeräusche geführt wird. Es erscheint offensichtlich, dass die schlechte Unterscheidung zwischen Sprache und Rauschen/Nebengeräuschen einen Grossteil des Verlustes der Spracherkennungsleistung ausmacht. Daher ist das sichere Unterscheiden zwischen Sprache und Rauschen/Nebengeräuschen ein grundlegendes Problem, dessen Untersuchung auf eine Verbesserung der Spracherkennungsleistung in stark verrauschten Komunikationssystemen zielt. Einige Studien haben zu einer Verbesserung der Unterscheidung von Sprache und Rauschen/Nebengeräuschen beigetragen und damit die Spracherkennungsleistung unter ungünstigen Bedingungen erhöht. Dieser Artikel schlägt Lösungen für das Problems der Unterscheidung von Sprache und Rauschen/Nebengeräuschen vor. Zunächst werden Vorverarbeitungen zur Spracherkennung betrachtet. Dazu werden zwei Versionen Rauschen/Nebengeräuschen, der auf statistischen Kriterien basiert, vorgestellt und verglichen. Letztlich wird eine Technik zum Filtern fälschlich Sprachsegmente vorgeführt.

Résumé

Les performances de la reconnaissance sont fortement dégradées lorsque les systèmes de reconnaissance sont employés sur des réseaux téléphoniques particulièrement difficiles et dans des environnements bruités. Il apparaît évident

* Corresponding author.
E-mail addresses: lamia.karray@francetelecom.com (L. Karray), arnaud.martin@univ-ubs.fr (A. Martin).

0167-6393/02/$ - see front matter
doi:10.1016/S0167-6393(02)00066-3

que la détection de parole/non-parole est une source importante de cette dégradation. Ainsi la robustesse de la détection de parole est un problème crucial à examiner pour améliorer les performances de la reconnaissance pour des communications très bruitées. De nombreuses études ont conduit à améliorer la robustesse de la détection de parole/non-parole pour une utilisation de la reconnaissance de parole dans des conditions difficiles. Ce papier propose des solutions pour l'amélioration de la détection de parole/non-parole en environnement très bruité. Des pré-traitements à la détection de parole sont d'abord considérés. Nous proposons et comparons ensuite deux versions d'un algorithme de détection de parole/non-parole, fondées sur des critères statistiques. Finalement, une technique de post-traitement est introduite dans le but de rejeter les détections de bruits prises pour de la parole.

Keywords: Speech/non-speech detection; Spectral subtraction; Adaptive algorithm; Likelihood ratio criterion; Wavelets

1. Introduction

Nowadays, interactive voice response systems are increasingly used, involving speaker-independent recognition of given vocabularies over the telephone network. In the same time, the recent rising crescendo of activity in mobile communication domain offers new opportunities for applications of speech recognition. However, the flexibility of such mobile networks offers the possibility to call from anywhere and at any time: calls can be made in various environments (e.g. indoor, outdoor, stopped car, running car). This results in very noisy speech.

In very noisy environments, the recognition performance degrades drastically. Robustness to noise is then required for an efficient use of the recognition systems especially in mobile networks context. Various studies have been conducted in this direction (Savoji, 1989; Mauuary and Monné, 1993; Junqua et al., 1994; Agaiby and Moir, 1997; Karray and Mauuary, 1997). Several pre-processing techniques have been developed in order to reduce the noise effects in the speech to be recognized. Enhancement procedures like spectral subtraction (Berouti et al., 1979; Mokbel et al., 1997) remove ambient noise. The transmission effects are reduced using equalization techniques such as cepstral normalization and adaptive filtering (Hermansky et al., 1993; Mokbel et al., 1995).

Moreover, high performance speech recognition requires efficient speech detection, especially in noisy environments. When using isolated-word recognition techniques, it is well known that a major cause of error in automatic speech recognition is inaccurate detection of the endpoints. Many speech/non-speech detection (SND) techniques are based on energy levels (Savoji, 1989). However, in real environments, the speech signal is corrupted by additive noise and energy mean based parameter may be insufficient for the correct detection of speech if the signal-to-noise ratio (SNR) is low.

Therefore, this paper provides some solutions aiming to improve the speech detection robustness in noisy wireless environments. The paper is organized as follows.

In Section 2, we describe the evaluation context: databases and modeling issues. The SND system is described in Section 3. This adaptative speech detection algorithm provides the starting point for investigating three different improvements, which are described in Sections 4–6. Conclusions are presented in Section 7.

2. Speech databases and modeling issues

Since this paper deals with SND in the observed signal, the considered databases contain continuously recorded speech. This means that the whole communication is continuously recorded, including words and also silence or noise between the words. Thus, we obtain what we called a continuous recording isolated words database.

Two databases are used for evaluation. One is recorded over public switched network (PSN). The other is a global system mobile (GSM) database.

2.1. PSN database

The PSN database is a 25 word database collected in field condition using an interactive voice response system giving movie programs. The obtained corpus contains about 30,500 hand segmented and labeled tokens, of which 76% are vocabulary words, 16% are out-of-vocabulary (OOV) words, and 8% are noise segments.

This database is used, in this paper, to evaluate the effect of the proposed solutions and check their compatibility in the context of relatively quiet environments (see Section 5.4).

2.2. GSM database

We use a laboratory GSM database of 51 words (digits and several command words) collected continuously, over the wireless GSM network.

Several call environments are considered, as follows:

- indoor: office, house, etc. (relatively quiet), with an average SNR of 17.5 dB,
- stopped car: (also relatively quiet, if the car windows are closed !), with an average SNR of 18.8 dB,
- outdoor: street, market, etc. (generally noisy with impulsive noises), with an average SNR of 17.1 dB,
- running car: with more or less speed and ambient noise (generally noisy, with a varying level of noise), with an average SNR of 16.6 dB.

The principal difference between the first two conditions and the last two conditions is the number of impulsive noises. Moreover outdoor and running car environments have an average SNR lower than indoor and stopped car environments.

About 500 labeled communications are used with almost the same proportion of each environment (26% indoor, 22% outdoor, 29% from stopped car and 23% from running car). The acquisition of the whole communications results in long regions without speech, and therefore in many regions of noise. Hence, in the obtained signal, not only are ambient noises frequent (especially in outdoor and running car calls), but the GSM transmission effects are also very disturbing. Therefore, different labels of noise and OOV words were added to the initial vocabulary words. This results in a database of 35,995 labeled segments, of which 64% are vocabulary words, 7% are OOV words and 29% are noise segments (16% background noises (BN), 9% GSM channel distortion (GSMN), and 4% remaining echoes).

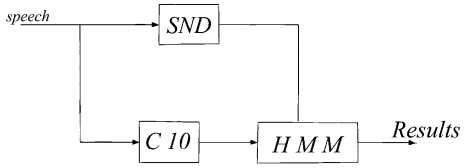

Fig. 1. Global recognition system. C10 is the acoustic analysis module and SND is the one of speech/non-speech detection.

2.3. Speech recognition system

The speech recognition system developed in France Télécom R&D is based on hidden Markov models (HMMs) and used in speaker-independent mode (Sorin et al., 1995). The feature vectors used in our experiments contain 27 coefficients. First, the energy on a logarithmic scale and the first 8 Mel frequency cepstrum coefficients are computed on 32 ms frames; with a frame shift of 16 ms. Then, first and second derivatives of these nine coefficient vectors are estimated on a 5-frame window.

Left–right HMMs with 30 states are used to model the vocabulary words, and silence models are placed on both sides of the vocabulary models to avoid precise detection of the words to recognize. A simple Gaussian probability density function with a diagonal covariance matrix is associated with each HMM state. The global system: acoustic analysis, SND and HMM modeling, is depicted in Fig. 1. The SND system is described in the next section.

3. Speech/non-speech detection

It was observed that a significant number of recognition errors are caused by non-efficient SND

(Mauuary, 1994), since some utterances may be truncated or even omitted. Field and wireless communications are very good illustrations of such problems. Therefore, several speech detection systems were studied and evaluated (Mauuary, 1994; Junqua et al., 1994). In this paper, an adaptive five state automaton is considered as a reference system to be improved.

The five states are: *silence*, *speech presumption*, *speech*, *plosive or silence* and *possible speech continuation* (Mauuary and Monné, 1993).

The transition from a given state to another one is conditioned by the frame energy and some duration constraints. These transitions between the different states determine the boundaries of speech segments. The speech presumption, plosive or silence and possible speech continuation states are introduced in order to cope with the energy variability in the observed speech and to avoid various kinds of noise. Hence, the speech presumption state avoids the automaton going in the speech state when the energy increase is due to an impulsive noise. The plosive or silence state takes the energy decrease within the speech (typically plosive) into account. The *possible speech continuation* state is indicative of the silence between two words within a group of words.

3.1. Baseline energy based adaptive detection algorithm

For adaptive detection, the energy requirements are based on an estimation of the SNR of the observed speech signal. The technique relies on a comparison between short-term and long-term estimates of the signal energy.

The short-term estimate is the mean energy computed over the last K frames, where K is the short-term span. As for the long-term energy, it is estimated recursively, when the automaton is in the silence state, as follows:

$$\mathrm{LTEE} \leftarrow \mathrm{LTEE} + (1-\lambda)(\mathrm{energy} - \mathrm{LTEE}),$$

where *LTEE* denotes the long-term energy estimate and λ is the forgetting factor (we use $\lambda = 0.99$). Then, the difference between short-term and long-term estimates is compared to a given threshold on the energy.

3.2. Evaluation procedure

It was shown (Mauuary, 1994) that some detection errors can be recovered by a rejection module used in the decoding process. For instance, a noise input can be rejected in the rejection module, which allows recovery from the speech detector error. Therefore, the detector evaluation procedure takes the whole recognition system into account. This evaluation is based upon the comparison between the reference and the recognized segments. The reference segments correspond to the hand segmentation and labeling of the calls. The recognized segments correspond to the automatic segmentation (by the speech detector) and labeling (by the recognition module) of the calls.

In practice, original signals from the continuous recording database are automatically segmented using the considered five state speech detector, then automatically labeled using a HMM. This HMM is trained on reference utterances extracted from the original signals after a hand segmentation and labeling. These automatically detected and labeled segments are then compared to the reference hand segmented and labeled ones, in order to evaluate the SND.

Substitution of vocabulary words and false acceptance of OOV words and noises are considered as major errors. Rejection of vocabulary words are less severe errors, but also important.

3.3. Results in adverse conditions

We usually give the evaluation results in terms of severe error rates (substitution errors and false acceptance errors) and false rejection error rates, as shown in Fig. 2. This figure illustrates the results obtained on the GSM database using the adaptive SNR-based endpoint detection algorithm. We give the results of a global evaluation, but we split false rejections into rejection errors due to the recognition model (false labeling of vocabulary word segments) and those resulting from a detection error (non-detection of vocabulary word segments). These results demonstrate the recognition and detection difficulties in such noisy conditions. Notice that the difficulties increase for outdoor and running car calls.

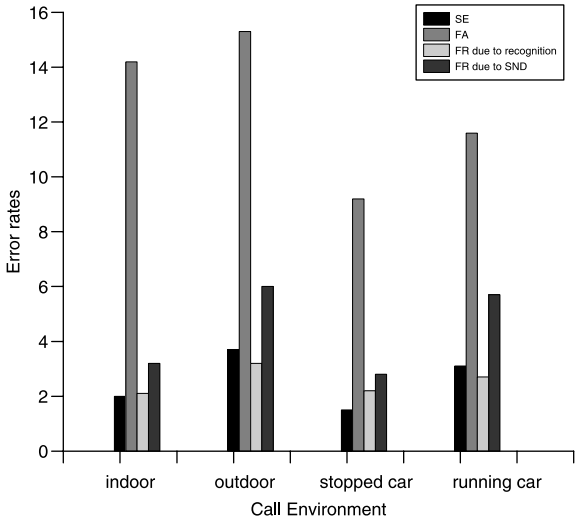

Fig. 2. Global evaluation of original speech recognition over the GSM network. Error rates are given in adverse call environments. Percentages of substitution errors (SE), false acceptances (FA) and false rejection (FR) are plotted. Distinction is made between FR due to the speech detector (SND) and those due to the recognition module.

In the following, we will propose some solutions in order to improve the poor recognition performance. Three aspects will be considered:

1. Pre-processing the observed speech in order to improve its quality prior detection.
2. Improving the SND algorithm within the SND system.
3. Post-processing the output detection in order to eliminate the wrongly detected noise segments.

4. Speech enhancement prior detection

Several pre-processing techniques may be used in order to reduce the channel effects of the telephone network and/or to enhance the speech signals in the presence of ambient noise (Mokbel et al., 1997). Two channel effect removing techniques (cepstral normalization and adaptive filtering (Mokbel et al., 1995, 1997)) and also spectral subtraction were shown to be efficient for the recognition module in the PSN environment. Here, we consider spectral subtraction and its effect on the whole recognizer including SND.

4.1. Spectral subtraction

Recall that spectral subtraction involves estimating the mean noise spectrum in the non-speech parts of the signal and subtracting this estimate from the frame spectra.

The noise spectral features considered in the algorithm are the mean and variance of the spectral densities in all the frequency bins, as well as the mean and variance of the filter-bank outputs.

To better understand the algorithm, we will consider the observed signal $x(t)$ which contains speech and noise,

$$x(t) = s(t) + n(t),$$

where $s(t)$ denotes clean speech signal and $n(t)$ an additive noise.

By supposing x, s and n centered distributions, and clean speech and noise decorrelated, we obtain in the spectral domain,

$$\Gamma_x(f) = \Gamma_s(f) + \Gamma_n(f).$$

For a given frame we observe the spectral density: $\Gamma_x(f)$. The corresponding spectral density of the clean speech can be estimated as follows:

$$\widehat{\Gamma}_s(f) = \Gamma_x(f) - \widehat{\Gamma}_n(f).$$

In practice, $\widehat{\Gamma}_n(f)$ is estimated in the non-speech parts of the observed signal. This estimation is updated in the silence periods, which are detected using some energy constraints. For more details see for example (Berouti et al., 1979).

Notice that this technique makes the assumption of stationary noise (at least for a word duration). This hypothesis is not true in the case of impulsive noise observed in the cellular network communications. Besides, a good estimation of noise spectrum requires a good detection of non-speech parts in the observed signal.

4.2. Evaluation

This pre-processing technique is applied on the GSM database. The enhanced continuously recorded database is segmented using the considered 5 state speech/non-speech detector. For the detection algorithm, we used the initial adaptive SNR-based algorithm.

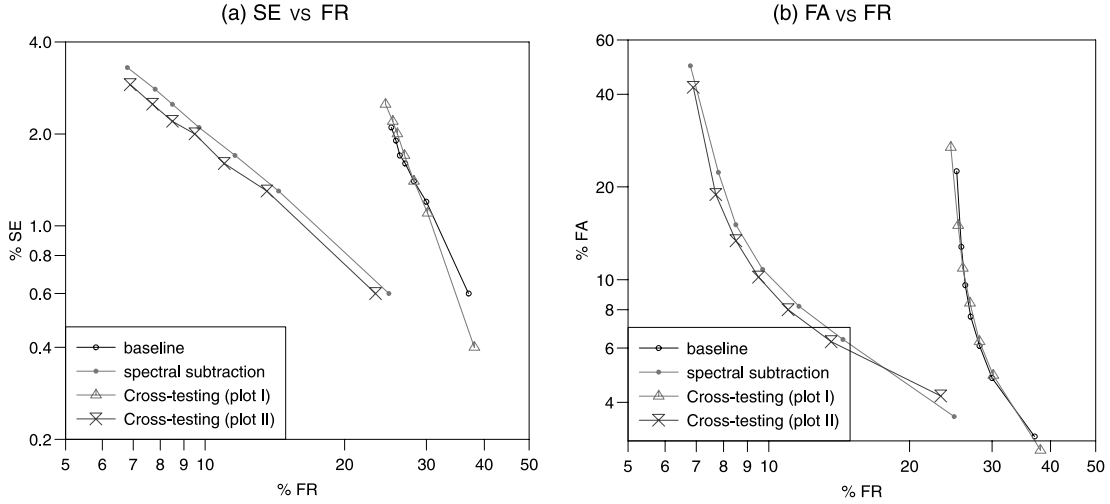

Fig. 3. Global evaluation results in GSM environment. We plot substitution error rates (%SE) and false acceptance rates (%FA) function of the false rejection rates (%FR). The "baseline" curve corresponds to the original signal (without pre-processing) and the "spectral subtraction" curve to the pre-processing speech prior segmentation and training. The plots I and II result from a cross-testing.

In order to take the recognition module into account, the training data (of the HMM model used for automatic labeling) is also pre-processed. Experimental results are summarized in Fig. 3.

The different points of the curves are obtained by varying the weight of the garbage models with respect to the weight of the vocabulary model. On the same figure, we summarize the results obtained, using the same procedure as above, with the same signals but after spectral subtraction pre-processing.

We notice that the application of spectral subtraction on the original speech improves the results considerably. Typically, for the same false rejection rate, the false alarm rate as well as the substitution errors are reduced. This decrease is a result of the improvement in the different recognition system modules. However, we will show, in the following, that it is mainly the SND module which was improved.

4.3. Investigations into the reasons for differences in performance

If we segment the original signal and only apply spectral subtraction after the speech detection stage (see plot I in Fig. 3), we notice that the results are almost the same as the baseline results. We also examined the case when we process only the signal to be segmented and use the original speech to build the model and to label the detected segments (this corresponds to plot II in Fig. 3). We notice that we achieve almost the same results as when the spectral subtraction was applied not only at the segmentation stage but also at the model training and labeling stages. Hence, only the enhancement of the signal to be segmented improves the results. This proves that the main improvement takes place in the detection module.

Another experiment shows that the improvement is due to the effect of the considered pre-processing on the signal energy. It consists in modifying the spectral subtraction algorithm in such a way that only the signal energy is processed. For the experiment the original speech is pre-processed in the same conditions as above but using the modified algorithm, which means that only its energy is pre-processed. Results are summarized in Fig. 4.

We can easily see that we have almost the same results as with the initial spectral subtraction pre-processing. Recall that the detection algorithm is

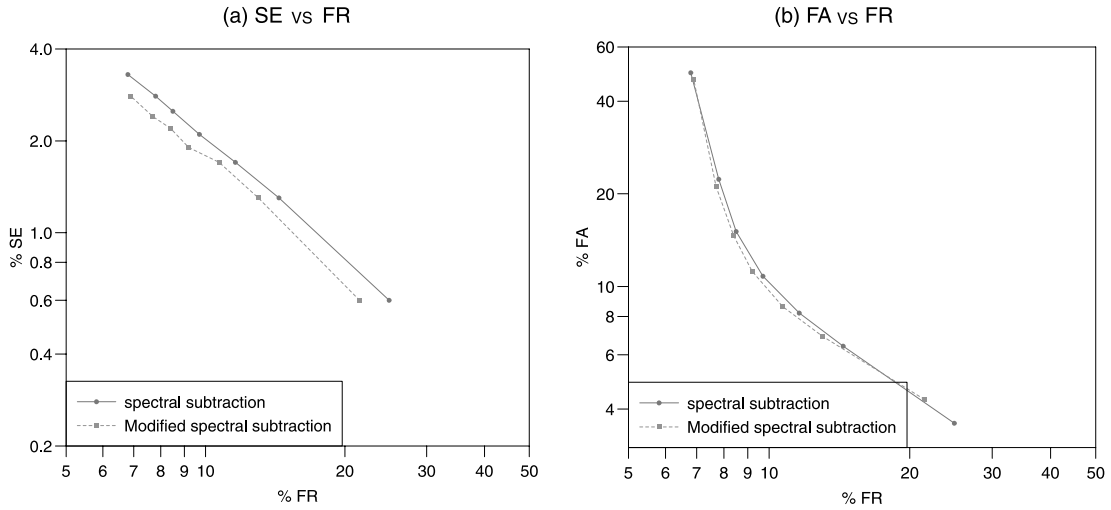

Fig. 4. Results when processing energy only. We plot substitution error rates (%SE) and false acceptance rates (%FA) as a function of the false rejection rates (%FR). The plots result from a global evaluation in GSM environment after pre-processing by the modified spectral subtraction algorithm (see text above).

based on the signal energy parameter alone. On the other hand, the contribution of this parameter is less important in the other modules of the whole recognition system (Mokbel et al., 1997). This provides the proof of spectral subtraction efficiency in speech detection, with only the signal energy pre-processed.

Hence, since spectral subtraction improves mainly the detection module, pre-processing the training data set is not necessary (see Fig. 3). Besides, only the signal energy needs to be filtered.

As a conclusion of these results, we notice that spectral subtraction reduces the error rates considerably. This could be easily explained by the fact that spectral subtraction allows an important reduction of the disturbing ambient noise. Thus, speech/noise detection is improved in GSM communications.

However, more efforts are needed in order to improve the robustness to GSM transmission effects. For instance, impulsive noise has an important contribution in the transmission distortion over the GSM channels. But, it could not be removed by the spectral subtraction technique recalled above. Therefore, special treatments will be investigated in order to perform robust recognition and speech detection for communications over the GSM network (see Section 6).

In the following, we will focus on the improvement of the endpoint detection algorithm robustness to noise by using other criteria for speech/non-speech distinction.

5. Detection algorithms based on statistical criteria

In this section, we propose two new criteria for the SND algorithm in order to improve recognition performance. The first criterion is based on noise statistics. The second one considers both noise and speech statistics.

5.1. Algorithm based on noise statistics

We consider the same automaton as the one described in Section 3. However, the transition between the 5 states (silence, speech presumption, speech, plosive or silence and possible speech continuation) is, in this case, based on a statistical criterion and duration constraints (the same as in the adaptive SNR-based algorithm).

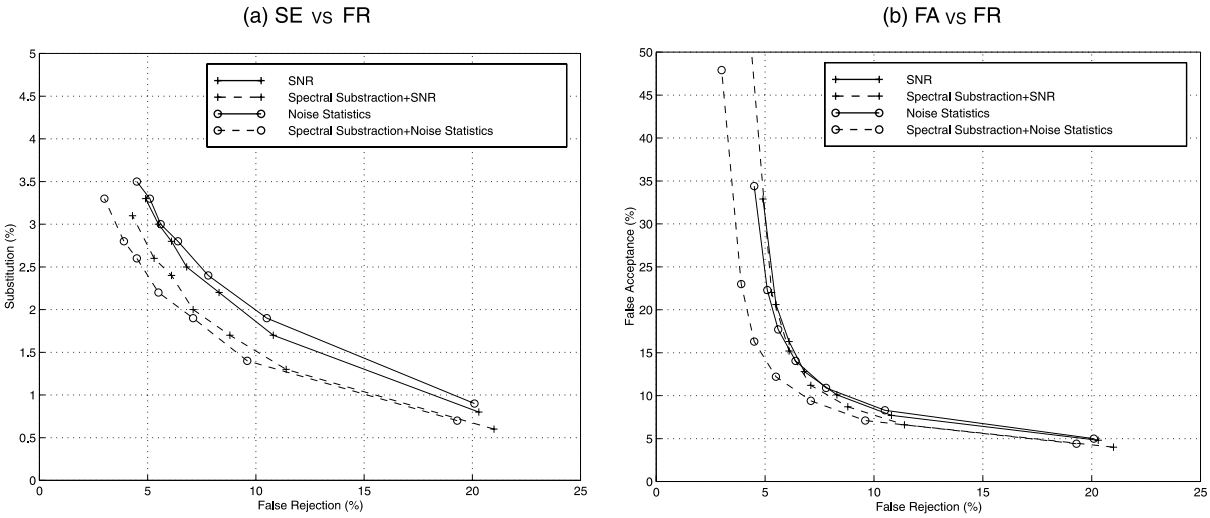

Fig. 5. Algorithm based on noise statistics. Evaluation results in GSM environment. We plot substitution error rates (%SE) and false acceptance rates (%FA) as a function of the false rejection rates (%FR). Results are given for SNR initial algorithm, the new statistical algorithm and their combination with spectral subtraction.

The idea involves testing the hypothesis of noise, for each observed frame. Therefore, we consider a normal distribution (μ, σ) for noise statistics. Then, for each frame, we compute the *critical ratio*,

$$r(x) = \frac{x - \mu}{\sigma},$$

where x is the current frame energy. This ratio is compared to a given threshold in order to decide if the considered frame belongs to noise or not.

The noise hypothesis is accepted for critical ratio values within a 95% confidence interval. The noise statistics are estimated recursively when the automaton is in the silence state as follows:

$$\mu \leftarrow \mu + (1 - \lambda)(x - \mu).$$

As for the variance estimation, we first estimate the second order moment μ_2,

$$\mu_2 \leftarrow \mu_2 + (1 - \lambda)(x^2 - \mu_2).$$

Then, the variance σ^2 is easily obtained,

$$\sigma^2 = \mu_2 - \mu^2.$$

Tested on the GSM database described above, the statistical criterion results in a slightly more robust algorithm compared to the initial one based on SNR estimation. The results are shown in Fig. 5.

However, recognition performance in noisy environments still requires more improvement. In order to increase robustness to noise, we try to reduce noise effects by pre-processing the observed speech signal prior to detection. Therefore, we use spectral subtraction, as this technique had previously been shown to be very efficient in such conditions (see Section 4). The results using spectral subtraction are included in Fig. 5. The pre-processing by spectral subtraction enhances the observed speech and increases the performance improvement.

5.2. Detection algorithm based on speech and noise statistics

In this version of the detection algorithm, we consider both noise and speech statistics. Notice that, in adverse conditions, the speech parts of the observed signal are corrupted by noise (ambient noise or transmission distortion, etc.). Hence, the speech statistics actually represent the statistics of speech plus noise.

Since the aim of SND is to distinguish between noise (or non-speech) and speech frames, we con-

sider two distributions: one for noise and one for speech. Then, we decide to which distribution each frame of the observed signal belongs.

In other words we have to deal with a hypothesis testing problem, with

- H_0: noise (or non-speech),
- H_1: speech + noise.

The decision rule considers the most probable hypothesis, according to the Bayesian approach. This results in a decision criterion based on maximum likelihood. Hence, for a given observed frame x, we compare the likelihood $\Pr(H_k/x)$ of the two hypotheses H_0 and H_1. Using Bayes formula and assuming the two hypotheses equally distributed, the problem is reduced to a comparison to 1 of the ratio: $r(x) = (P(x/H_0))/(P(x/H_1))$.

Hence, we end up with a likelihood ratio criterion.

The decision rule is the following:

- if $r(x) > 1$ the frame x belongs to a noise (or non-speech) segment (hyp. H_0),
- if $r(x) \leqslant 1$ the frame x is a speech frame (hyp. H_1).

The distributions corresponding to H_0 and H_1 are determined recursively by estimating their means and variances as described in Section 5.1. Noise statistics are updated every time the automaton is in the silence state. Speech statistics are updated every time the automaton is in the speech state. For recursive adaptation, we use a forgetting factor of 0.99 for noise statistics, and 0.95 for speech statistics, assuming that noise is more stationary than speech.

Tested on the GSM database, this extended statistical approach results in a more robust algorithm compared to the initial one based on SNR estimation and the one based on noise statistics only.

Recognition results are evaluated using the different algorithms mentioned above. The new algorithm performance is then compared to the previous ones. In Fig. 6, we summarize the results obtained with the different algorithms.

This figure shows that the algorithm based on speech and noise statistics improves the overall recognition performances, especially when combined with spectral subtraction.

Moreover, the overall measured decrease in the error rate actually depends on how noisy the observed signal is. In the following, we will provide a

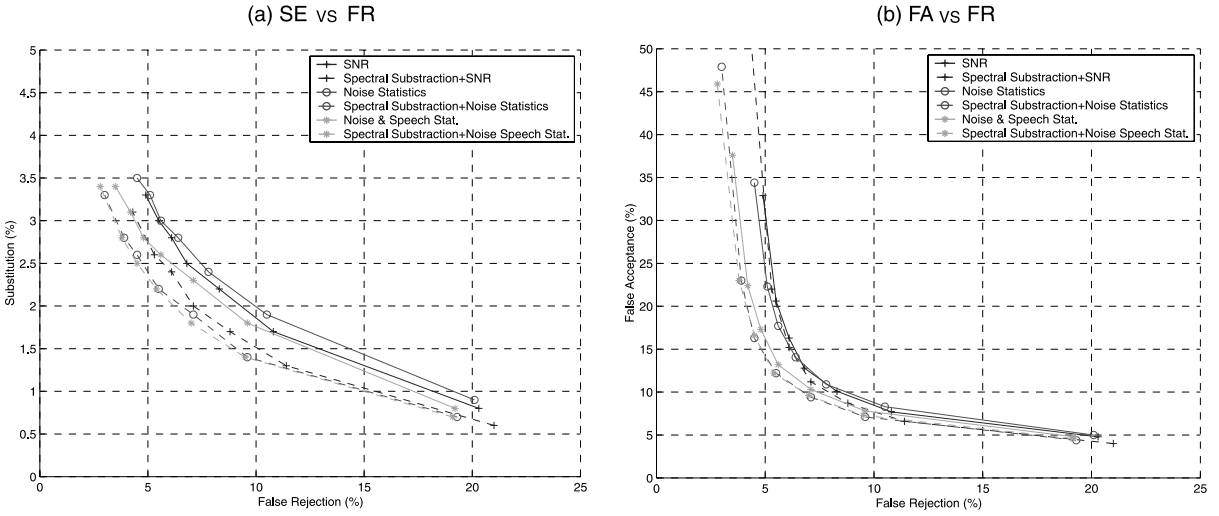

Fig. 6. Algorithm based on noise and speech statistics. Evaluation results in GSM environment. We plot substitution error rates (%SE) and false acceptance rates (%FA) as a function of the false rejection rates (%FR). Results are given for SNR initial algorithm, the noise statistical algorithms and their combination with spectral subtraction.

detailed study of the different algorithms' behavior in adverse call environments (indoor, outdoor, stopped car and running car).

5.3. Performance in adverse call environments

The GSM database used for the experiments contains calls from several environments. Indoor and stopped car conditions are generally relatively quiet. But the other difficult environments (outdoor and running car) can be very noisy, and usually present very high acoustical variations.

The results obtained with the different SND algorithms described above (based on SNR, noise statistics or noise and speech statistics) are given, in Fig. 7, separately for each condition.

We notice different behaviors according to the call environment. Hence, we obtain more improvement in noisy environments than in quiet ones. This could be easily explained by the fact that quiet communications contain less noise and less acoustical variations than difficult conditions. For noisy environments, the estimation of noise and speech statistics increases the detector ro-

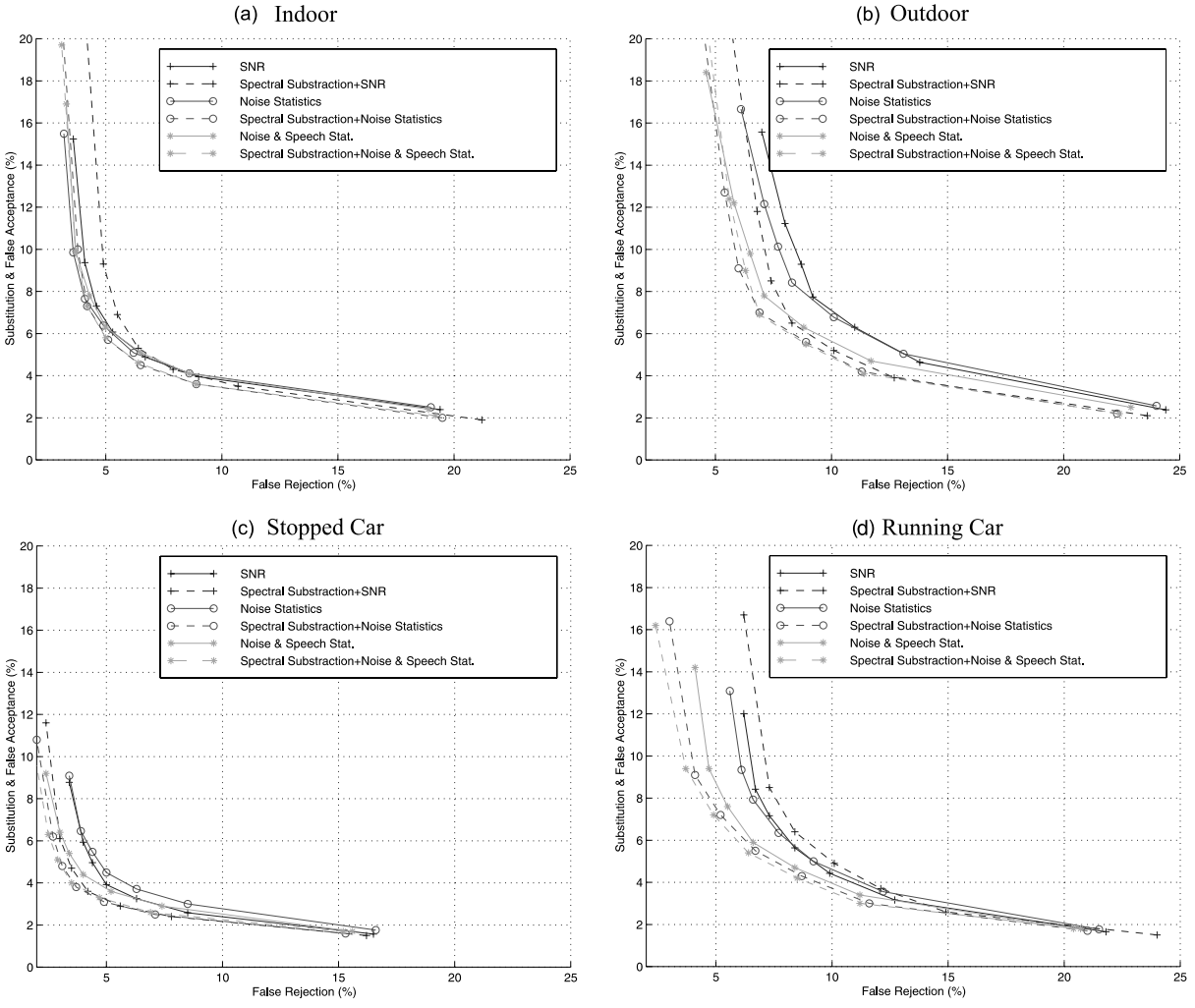

Fig. 7. Evaluation of the different SND algorithm in adverse call environments. For each call environment, we plotted the severe error rates (false acceptance and substitutions), as a function of the false rejection error rates. We also show the effect of spectral subtraction combined with the different algorithms, in the different call environments.

bustness to the variations of the ambient noise characteristics (for instance, due to speed variations in the running car case). We also notice that spectral subtraction pre-processing does not improve the results for either of the statistics-based speech-detection methods when tested in the relatively quiet environments. This finding is also easily explained by the lack of additive noise in these environments, so spectral subtraction could not improve the performance. However, in difficult (outdoor and running car) conditions, the effect of the pre-processing is more noticeable.

5.4. Consistency in PSN environment

In order to check the compatibility of the proposed algorithm, the same techniques are tested for speech recognition over the PSN network. The PSN continuously recorded field database, described in Section 2.1, is used. The results are shown in Fig. 8.

The performances with the different solutions do not differ significantly, presumably because the PSN speech contains little ambient noise. We cannot objectively compare the performances on the PSN database and on the GSM database, since one is a field database (PSN case) and the other is a laboratory database (GSM case). However, it is obvious that the major difference is the ambient noise and the signal variability.

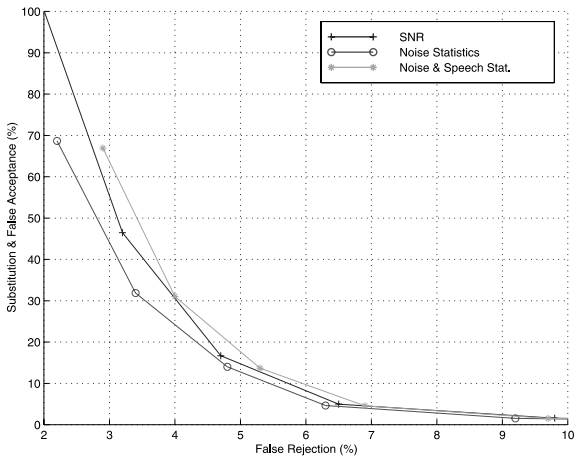

Fig. 8. Global evaluation of original speech recognition over the PSN network. Error rates are given for the different algorithms. We plot severe error rates (substitutions and false acceptance) as a function of the false rejection rates.

6. Post-detection denoising technique

Despite the improvements, many segments of noise may be wrongly detected by the speech/non-speech detector, which increases the false acceptance errors. These errors are due to the poor quality of the observed signal. In the GSM speech signal one can notice several noises: ambient noise (people in the street, cars, etc.), impulsive noise (due to the cellular network transmission), etc. However, a spectral study of the GSM signal shows that the various kinds of noise are not located in the same energy band. Therefore, we introduce in this section a post-processing technique based on the localization of different kinds of noise in different sub-bands. The algorithm is based on a denoising technique using a discrete wavelet transform of the detector's output segments.

6.1. Wavelet transform and denoising

Several previous studies have shown thresholding in the wavelet domain to be an effective technique in denoising (Donoho, 1995; Burley and Darnell, 1997; Burstein and Evans, 1997; Downie and Silverman, 1998). We will not provide here the derivation or a detailed discussion of the wavelet transform, more details and discussions could be found in (Vetterli and Kovacevic, 1995). Wavelet based noise reduction takes advantage of the wavelet transform simultaneous localization of time and frequency information. In the wavelet domain, scale corresponds to frequency. Coarse scale wavelets are localized in frequency, while fine scale wavelets are localized in time. The advantage of this localization is that modification can be made to the signal at particular scales without affecting, noticeably, the remainder of the signal for all time. This is in sharp contrast to filtering in the Fourier transform domain. Therefore, application of wavelets to signal processing has a great interest. Due to localization properties, and the

energy preserving nature of the wavelet transform, the signal will be represented in the wavelet domain predominately by a small number of large coefficients corresponding to the time-scale location of the signal phenomenon.

In this section we investigate a denoising method based on a reduction or suppression of the contribution of noise while reconstructing the initial transformed segment.

An example of speech and noise in a portion of the observed signal is shown in Fig. 9. It contains an example of GSM noise followed by a vocabulary word, in a (relatively) clean environment. This portion of signal is filtered using discrete wavelet transform. Fig. 10 presents the obtained sub-bands (seven sub-bands). This figure illustrates the possibility to localize speech and noise in the sub-bands.

The idea involves taking advantage of the time–frequency localization properties of the wavelet transform, to reduce or suppress the contribution of the sub-bands where noise is dominating. Therefore, we need to localize the different kinds of noise and to distinguish them from the speech segments, in the different decomposition levels. We focus on two kinds of noise: GSMN and BNs that we would like to reject.

In order to localize them in the different decomposition levels, a statistical study of the energy in the different sub-bands is conducted, based on the time–frequency localization properties of the wavelet transform. For this purpose, we localize, for each segment, the maximum of energy M_1 in the first half of decomposition levels and M_2 the maximum in the second half. For example, if we consider 12 decomposition levels, M_1 is the maximum of energy in levels 1–6 (high frequencies), and M_2 the maximum of energy in levels 7–12 (lower frequencies). Figs. 11 and 12 illustrate the distribution of energy in the sub-bands for a GSMN segment (Fig. 11) and a real speech segment (Fig. 12).

Then, we compute the ratio: $R = M_1/M_2$.

A statistical study of this ratio shows that:

- For 95% of speech segments, we notice that $R > 5$, and only 0.5% of speech segments have $R < 2$;
- For 70% of GSMN segments and 50% of BNs segments, we notice that $R < 5$.

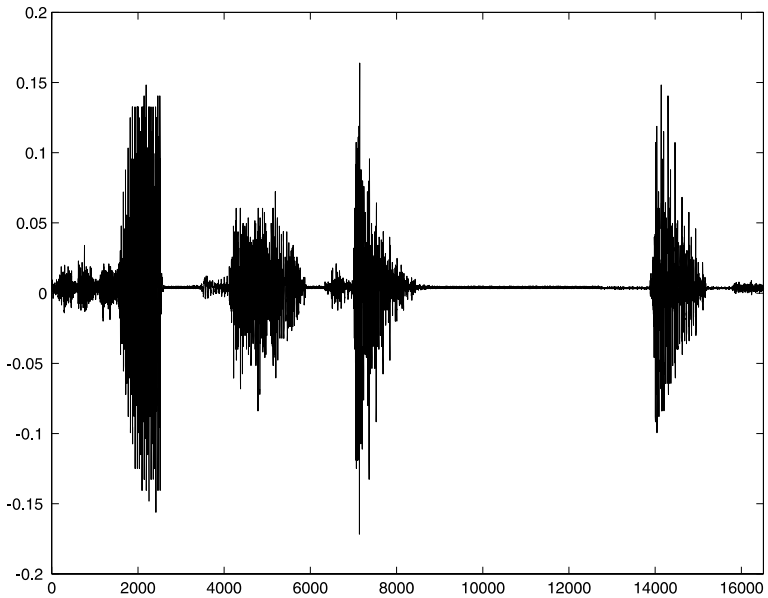

Fig. 9. Example of segment containing GSM noise (first peak) and speach (the three following peaks).

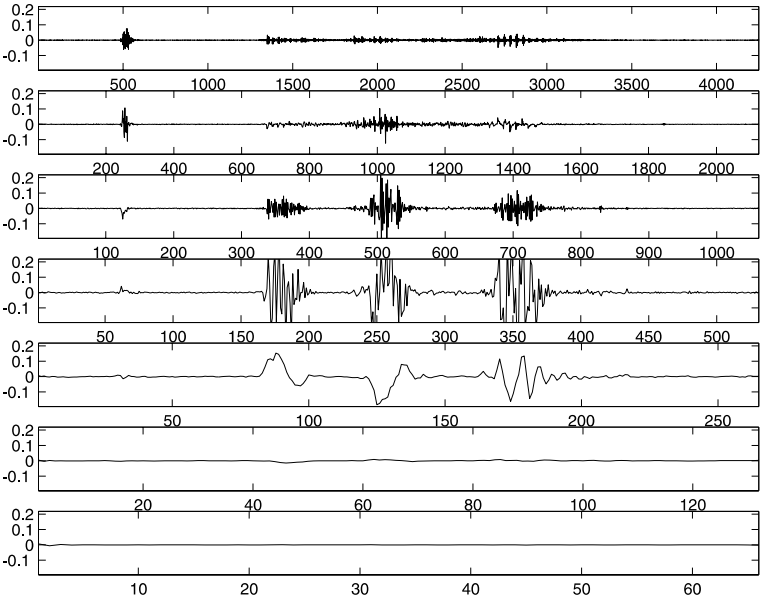

Fig. 10. Discrete wavelet decomposition of the signal portion of the segment depicted in the previous figure. Seven decomposition levels are considered using a 10-tap Daubechies filter.

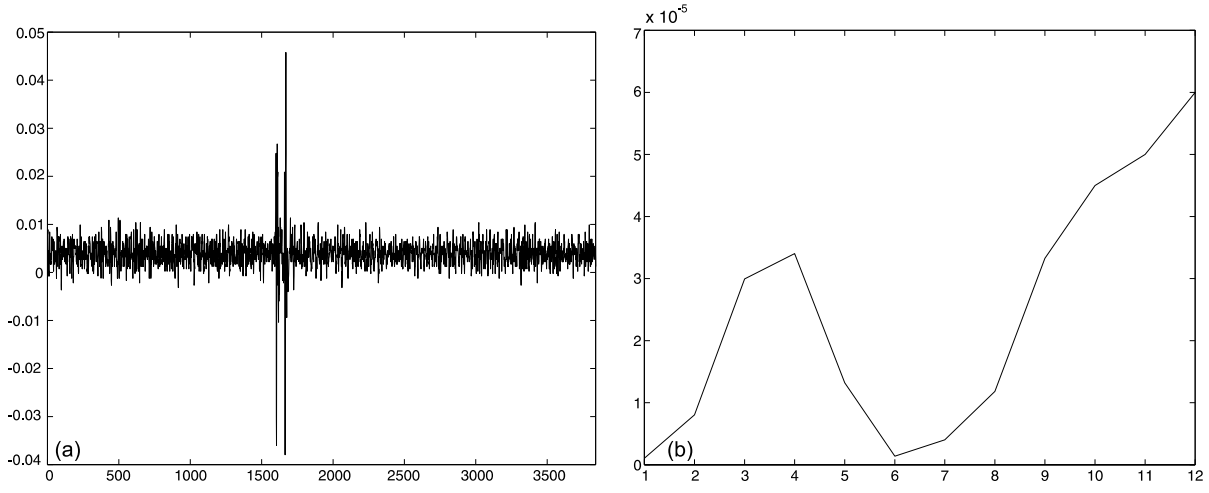

Fig. 11. (a) A GSM noise segment and (b) the corresponding distribution of the energy in the sub-bands. M_1 (first energy maxima) is reached in level 4, and M_2 (second energy maxima) in level 12. M_1 is slightly lower than M_2.

This study allows us to define a certain scale to measure the contribution of noise in the sub-bands. Hence, we end up with a discrimination criterion between speech and several kinds of noises. This leads to the following decisions:

1. If $R > 5$, the segment is more likely to be speech, we keep it;
2. If $2 < R < 5$, the segment may be corrupted by noise, we can denoise it if we reduce the contribution of the first decomposition levels (high

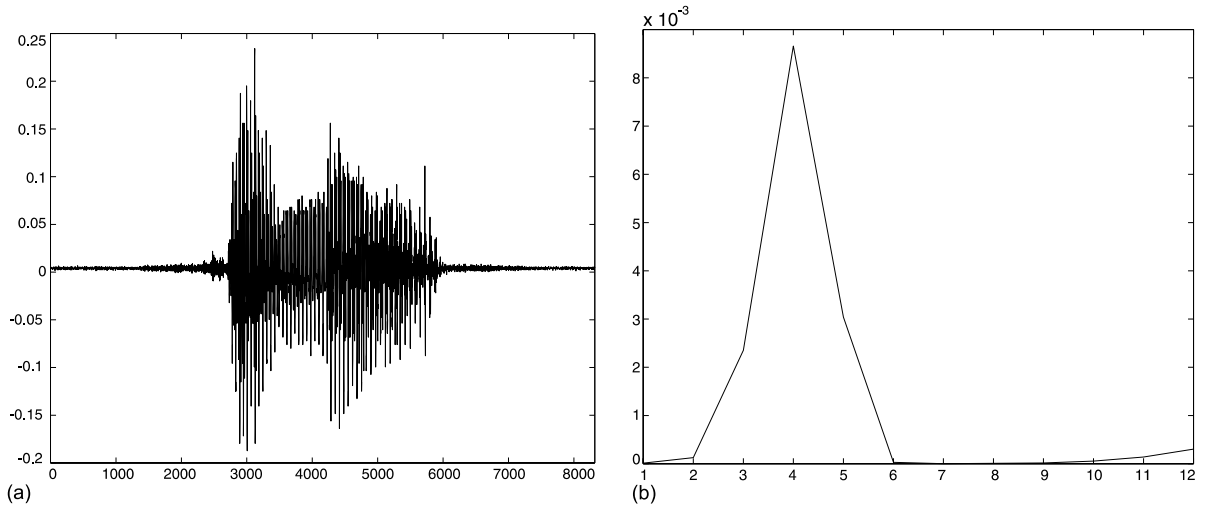

Fig. 12. (a) A speech segment and (b) the corresponding distribution of the energy in the sub-bands. M_1 is reached in level 4, and M_2 in level 12. M_1 is much higher than M_2.

frequencies) by under-weighting them while reconstructing the segment;

3. If $R < 2$, the segment is more likely to be noise (GSMN), we reject it;

This criterion is then used as a post-processing of the basic endpoint detection. This post-processing results in a reduction or suppression of the contribution of the sub-bands where noise is dominating. Thus, several noisy segments could be rejected, and corrupted speech could be enhanced, before the recognition procedure.

6.2. Experimental results

We apply the detection algorithm based on noise statistics, and the proposed denoising technique in the context of the GSM database.

The SND system is applied to this data, using the noise statistical criterion based detection algorithm. Then the proposed denoising technique is applied to the SND outputs, in order to reject the wrongly detected segments of noise, and to enhance the corrupted speech. For wavelet transform, we use a 10-tap Daubechies filter (Daubechies, 1988), with a decomposition depth of 12 (i.e., 12 decomposition levels).

The method, described above, is evaluated in terms of the percentage reduction of detection errors. A complete evaluation could be performed in terms of recognition performance, in order to quantify the enhancement effect.

Despite the robustness of the SND algorithm, the output of the SND contains some remaining non-speech segments. In our example, using the algorithm based on noise statistics, 14.5% of the detected segments are non-speech. In particular, 25% of them are due to GSMN and 18.6% are BNs. The post-processing technique, introduced in this paper, allows a significant reduction of non-speech wrongly detected segments (particularly, GSMN and BNs), as it is shown in Table 1.

Table 1
Evaluation results in GSM environment

Detected segments	Non-speech (%)	GSMN (%)	BN (%)	Speech (%)
Indoor	42	71	15	0.5
Outdoor	34	67	16	0.4
Stopped Car	28	76	5	0.6
Running Car	28	48	16	0.2

We give the number of segments rejected by the denoising post-processing of the detector's output. We also give the corresponding reductions with respect to the initial results of the SND system.

This table shows that the proposed denoising technique results in a reduction of 46% of non-speech wrongly detected segments (particularly, 69% reduction of GSMN and 9% reduction of BNs). However, some speech segments were also rejected. The analysis of these misrejected speech segments reveals that the corresponding speech is highly corrupted by noise. So, they are very unlikely to be correctly recognized. Hence, this misrejection would not decrease the overall recognition performances.

In the following, we will study the behavior of this denoising technique in different call environments.

6.3. Results in adverse call environment

We have shown above (Section 5.3) that the considered SND system has different performance according to call environment. Hence, we obtain more or less wrongly detected segments of noise. Consequently, the proposed post-processing could have different behavior according to the call environment. The results obtained with this technique applied as a post-processing of the SND algorithm mentioned above (based on noise statistics) are given in Table 2, separately for each condition. The results are given in terms of percentage of rejected segments (non-speech, GSMN, BNs and speech).

From Table 2, we notice the important rejection rates of the non-speech wrongly detected segments. However, these rates are different according to the call environment and to the kind of noise. Hence, the rejection rates for GSMN are bigger than for BNs, and we reject more GSMN in quiet environments (indoor and stopped car) than in noisy environments (outdoor and running car). This is due to the fact that a detected segment containing GSMN (that contains also some frames before and after the noise itself) in a quiet communication is less corupted by the BN. So, the efficiency of the rejection procedure is more important for GSMN, particulary in quiet environment.

Table 2
Evaluation results in several call environments

Segments	Non-speech	GSMN	BN	Speech
Number of detected segments	2545	882	656	2100
Number of rejected segments	1167	603	59	98
Reduction (%)	46	69	9	0.5

We give the reductions with respect to the initial results of the SND system.

7. Conclusion

In order to improve the performance of speech recognition systems, this paper dealt with the SND robustness to noise. Several solutions were proposed.

First, pre-processing techniques were considered. Spectral subtraction was shown to reduce the effect of noise in GSM communications, which improved the speech detection, and, consequently, the global recognizer performance.

Then, two detection algorithms based on statistical criteria were introduced. One using noise statistics, and the other is based on noise and speech statistics.

Finally, a post-processing technique using a wavelet based denoising was applied on the obtained detected segments. It resulted in a reduction of almost 50% of wrongly detected non-speech segments.

The different proposed solutions were evaluated in adverse call conditions over the cellular GSM network. The different SND algorithms were also evaluated in the PSN context, in order to check their consistency.

In conclusion, the different proposed solutions increase more or less the SND performance according to the call environment. Important improvements are noticed in noisy environments, such as outdoor or running cars. Incorporating a pre-processing technique like spectral subtraction enhances the improvements. Also, the application of the post-processing technique introduced in this paper allows the major portion of the wrongly detected segments to be rejected, so improving the final detection results.

Acknowledgements

The authors are grateful to Denis Jouvet, as well as the anonymous reviewers for their very helpful comments and criticisms that contributed to the improvement of this paper.

References

Agaiby, H., Moir, T.J., 1997. Knowing the wheat from the weeds in noisy speech. In: European Conference on Speech Communication and Technology, Greece, pp. 1119–1122.

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 208–211.

Burley, S., Darnell, M., 1997. Robust impulsive noise suppression using adaptive wavelet denoising. In: International Conference on Acoustics, Speech, and Signal Processing, April 1997, pp. 83–86.

Burstein, E., Evans, W., 1997. Wavelet based noise reduction for speech recognition. In: Robust Speech Recognition for Unknown Communication Channels, France, April 1997, pp. 111–114.

Daubechies, I., 1988. Orthonormal bases of compactly supported wavelets. Comm. Pure Appl. Math. 41 (7), 909–1294.

Donoho, D., 1995. Denoising by soft thresholding. IEEE Trans. Inform. Theor. 41, 613–627.

Downie, T.R., Silverman, B.W., 1998. The discrete multiple wavelet transform and thresholding methods. IEEE Trans. Signal Process. 46 (9), 2558–2561.

Hermansky, H., Morgan, N., Hirsch, H.G., 1993. Recognition of speech in additive and convolutional noise based on RASTA spectral processing. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 83–86.

Junqua, J.-C., Mak, B., Reaves, B., 1994. A robust algorithm for word boundary detection in the presence of noise. IEEE Trans. Speech Audio Process. 2 (3), 406–412.

Karray, L., Mauuary, L., 1997. Improving speech detection for wireless speech recognition. In: IEEE Workshop on Automatic Speech Rocognition and Understanding, Santa Barbara, December 1997, pp. 428–435.

Mauuary, L., 1994. Improving the performances of interactive voice response services. Ph.D. Thesis, Université de Rennes, Rennes (in French).

Mauuary, L., Monné, J., 1993. Speech/non-speech detection for voice response systems. In: European Conference on Speech Communication and Technology, Berlin, September 1993, pp. 1097–1100.

Mokbel, C., Jouvet, D., Monné, J., 1995. Blind equalization using adaptive filtering for improving speech recognition over telephone. In: European Conference on Speech Communication and Technology, pp. 141–1990.

Mokbel, C., Mauuary, L., Karray, L., Jouvet, D., Monné, J., Simonin, J., Bartkova, K., 1997. Towards improving ASR robustness for PSN and GSM telephone applications. Speech Communication 23 (1–2), 141–159.

Savoji, M.H., 1989. A robust algorithm for accurate endpointing of speech signals. Speech Communication 8, 45–60.

Sorin, C., Jouvet, D., Gagnoulet, C., Dubois, D., Sadek, D., Toularhoat, 1995. Operational and experimental French telecommunications services using CNET speech recognition and text-to-speech synthesis. Speech Communication 17 (3), 273–286.

Vetterli, M., Kovacevic, J.S., 1995. Wavelets and Sub-Band Coding. Prentice Hall, Englewood Cliffs, NJ.