Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

ELSEVIER

Speech Communication 48 (2006) 191–206

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Robust speech/non-speech detection based on LDA-derived parameter and voicing parameter for speech recognition in noisy environments

Arnaud Martin [a,*], Laurent Mauuary [b,1]

[a] ENSIETA/E^3I^2, EA3876, 2 rue F. Verny, 29806 Brest Cedex 9, France
[b] France Télécom R&D, 2 av.P. Marzin, 22307 Lannion Cedex, France

Received 25 October 2004; received in revised form 15 April 2005; accepted 20 July 2005

Abstract

Every speech recognition system contains a speech/non-speech detection stage. Detected speech sequences are only passed through the speech recognition stage later on. In a very noisy environment, the noise detection stage is generally responsible for most of the recognition errors. Indeed, many detected noisy periods can be recognized as a vocabulary word. This manuscript provides solutions to improve the performance of a speech/non-speech detection system in very noisy environment (for both stationary and short-time energetic noise), with an application to the France Télécom system.

The improvement we propose are threefold. First, noise reduction is considered in order to reduce stationary noise effects on the speech detection system. Then, in order to decrease detections of noise characterized by brief duration and high energy, two new versions of the speech/non-speech detection stage are proposed. On the one hand, a linear discriminate analysis algorithm applied to the Mel frequency cepstrum coefficients is incorporated in the speech/non-speech detection algorithm. On the other hand, the use of a voicing parameter is introduced in the speech/non-speech detection in order to reduce the probability of false noise detections.

Keywords: Speech/non-speech detection; Speech recognition; Noise reduction; LDA; Voicing parameter

1. Introduction

Spoken language is may be the most natural means of communication for us humans, which accounts for the increasing research on human–machine communication using voice. Interactive,

[*] Corresponding author. Tel.: +33 2 9834 8884; fax: +33 2 9834 8750.
E-mail addresses: arnaud.martin@ensieta.fr (A. Martin), laurent.mauuary@francetelecom.com (L. Mauuary).
[1] Tel.: +33 2 9605 1654; fax: +33 2 9605 3530.

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.07.005

voice-based systems have found many applications, especially in the booming mobile communication sector, that range from recognizing the name of someone one wishes to call, to entirely automatic information systems. Mobile phones allow people to use interactive voice response systems from anywhere and at any time; in particular, they will not only call from a quiet home or office, but also from very noisy environments such as urban transports or airports. However, in a very noisy environment, the performance of speech recognition systems decreases drastically. Therefore, robustness to noise is required for effective use of these systems, and this will especially be the case for calls made on a mobile phone.

In order to reduce the effects of noise, several enhancement procedures have been developed, such as cepstral normalization and adaptive filtering (Mokbel et al., 1997) or spectral subtraction (Karray and Martin, 2003). These techniques are shown to be efficient for the recognition of speech in a quiet environment. When evaluated on speech containing noise characterized by brief duration and high energy, these techniques are not effective anymore with respect to the error rate (Karray and Martin, 2003).

This degradation is in part due to the imperfect detection of sequences of speech amid sequences of noise. Indeed, in a very noisy environment, speech/non-speech detection systems tend to detect too many periods of noise, which entails errors in the automatic speech recognition stage. Therefore, having effective speech/non-speech detection is crucial. Several studies have been conducted to enhance the speech/non-speech detection system (e.g. in Savoji, 1989; Mauuary and Monné, 1993; Junqua et al., 1994; Huang and Yang, 2000; Martin et al., 2001; Karray and Martin, 2003). Many speech/non-speech detection techniques are only based on energy levels, e.g. in (Savoji, 1989; Mauuary, 1994; Martin et al., 2000). However, in a very noisy environment, noise can be characterized by a high energy. Thus, when the signal-to-noise ratio (SNR) is low, the use of additional parameters can produce better performances.

In this manuscript, we describe some solutions to improve the speech/non-speech detection robustness in very noisy environments for both stationary noise, and noise characterized by brief duration and high energy. We organize our paper as follows.

In Section 2, the evaluation context is described: first the used database and a speech recognition system are reviewed; next the evaluation procedure is described. The speech/non-speech detection stage with the three criteria given before is recalled in Section 3. Next, three different new improvements are presented. First a comparative study of the three criteria using a noise reduction system is made in Section 4. Afterwards, Section 5 presents a new method for speech/non-speech detection using a linear discriminate analysis (LDA) applied to Mel frequency cepstrum coefficients (MFCCs). In Section 6, a new speech/non-speech detection using a voicing parameter combined with the energy is studied. Lastly, conclusions are presented in Section 7.

2. Evaluation context

We describe here our speech database, the speech recognition system and the evaluation procedure used in our work. All the evaluations depend on both the database and the speech recognition system. For this reason, a precise evaluation procedure is defined. Several databases are used, some for the learning of the speech/non-speech detection parameters and some for experiments. Only one speech recognition system is used here. The learning database for the speech recognition model is not presented.

2.1. Speech databases

All databases contain continuously recorded speech. These databases are used for recognition of words belonging to the French language. The communication is recorded in its entirety, including words as well as silence periods, noise periods among words or noisy words. The databases were manually segmented by noting boundaries and labeling words (vocabulary and out-of-vocabulary) and also the different kinds of noises in the recordings. The segmentation was performed at the level of words and noises. With the aim of studying the effects of noise on the speech/non-speech detection system, databases have been separated into two almost equal-sized parts with respect to the SNR.

2.1.1. Learning databases

Two learning databases are used. One was recorded over the Public Switched Network (PSN). The other one is a Global System Mobile (GSM) database.

The first one, referred to as PSN_L, includes 1000 phone calls to an interactive voice response service giving movie programs. The corpus contains 25 different vocabulary words. This database was used for the learning of the three criteria thresholds of the speech/non-speech detection system presented in Section 3, and for the new criterion threshold presented in Section 5. The second learning database is referred to as GSM_L. The corpus of this laboratory database contains 51 vocabulary words, including 390 phone calls. This database was divided into two parts with respect to the SNR level in the recordings: less and greater than 18 dB. It was used for the same learning as the PSN_L, and for the comparative study in Section 4.

Manual segmentation on both learning databases gives 63% of vocabulary word segments, 9% of out-of-vocabulary word segments and 28% of noise segments on a total of 46063 labeled segments.

2.1.2. The test database

The test database was recorded over GSM and will be referred to as the GSM database. The corpus contains 65 vocabulary words repeated by different speakers. The 390 phone calls came from different environments: indoor, outdoor, stopped car and running car. This database is also divided into two parts according to the SNR: less and greater than 18 dB. The part with the SNR less than 18 dB contains many brief duration and high-energy noises and the background noise can be weak. Manual segmentation gives 85% of vocabulary word segments, 3% of out-of-vocabulary word segments and 11% of noise segments with a total of 29592 labeled segments.

2.2. Speech recognition system

The speech recognition system used in this paper is based on Hidden Markov Models (HMMs) and was developed by France Telecom R&D in (Mokbel et al., 1997). Recognition experiments were conducted using a feature vector containing 27 coefficients. First, the log energy, and the first eight MFCCs are computed on 32 ms frames; with a frame shift of 16 ms. Then, first and second derivatives of these nine coefficients are estimated. Left-right HMMs with 30 states are used to model the vocabulary words, and silence models are placed on both sides of the vocabulary word models to avoid imprecise detection that could cut the beginning or the ending of the words.

The whole system including acoustic analysis, speech/non-speech detection system, and speech recognition system is depicted in Fig. 1. The speech/non-speech detection system and speech recognition system work together. Therefore, the evaluation procedure of the speech/non-speech detection takes into account this interaction between both systems.

2.3. Evaluation procedure

It was shown in (Mauuary, 1994; Martin, 2001; Karray and Martin, 2003) that some detection

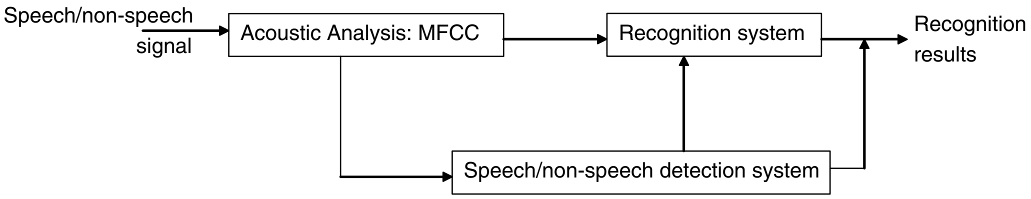


Fig. 1. Structure of the whole system.

errors like detection of non-speech periods (i.e. silence or noise detections) can be recovered by a rejection model used in the decoding process of the speech recognition system. Therefore, the speech/non-speech detection system evaluation procedure must take 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 those obtained through manual segmentation and labeling of the calls. The recognized segments correspond to those delivered by the automatic segmentation (by the speech/non-speech detection system) and labeling (by the recognition system) of the calls.

To evaluate the detection system, we consider two steps. Firstly automatic speech segment detection is compared to reference segments. Four different kinds of errors are considered:

- omission: a vocabulary word or an out-of-vocabulary word is not detected,
- insertion: a non-speech segment is detected, like speech,
- regrouping: several words separated by silence are detected as only one,
- fragmentation: one word is detected as several.

The rejection model of the recognition system can reject non-speech detections. These errors are called *recoverable errors*. Omission, regrouping and fragmentation errors unavoidably produce recognition errors. These errors are called *definitive errors*. Recoverable and definitive error rates are calculated with respect to the total number of manual speech segments (vocabulary and out-of-vocabulary words). For comparative studies, definitive errors vs. recoverable errors are plotted for different *adjusting thresholds* of the speech/non-speech detection system.

Secondly, the evaluation of the speech/non-speech detection is done through the recognition system evaluation. The recognition evaluation is achieved with the speech/non-speech detection. Three errors are considered:

- substitution: a vocabulary word is recognized as another vocabulary word,
- false acceptance: a non-speech period or out-of-vocabulary word is recognized as a vocabulary word,
- false rejection: a vocabulary word is rejected.

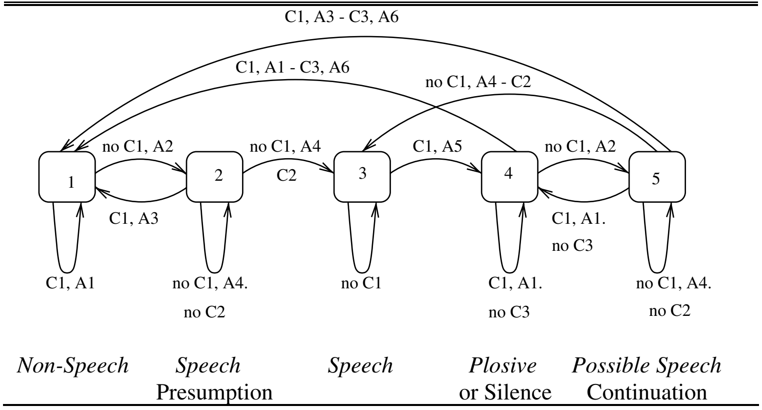
The false rejection error rate is calculated with respect to the vocabulary word manual segments, and substitution and false acceptance error rates are calculated with respect to the total number of manual segments. For comparative studies, a substitution and false acceptance error rate according to the false rejection error rate is plotted for different thresholds of the rejection model.

In order to calculate the statistical significance of the difference between two error rates, the confidence interval of the first error rate is calculated at 95% under Gaussian assumption of the rate. If the second error rate is outside the confidence interval of the first one, we consider that the difference is statistically significant.

3. The speech/non-speech detection system

In this paper, the adaptive five-state automaton shown in Fig. 2 is considered as the reference system to be improved. The five states are: *Non-Speech, Speech Presumption, Speech, Plosive or Silence* and *Possible Speech Continuation.*

The transition from one given state to another is controlled only by the energy contained in a frame and some duration constraints. Estimations for long-term and short-term signal energy are compared to an adaptive threshold called the *threshold energy*. This comparison, and the duration constraints, both determine the endpoint of the detection system. The *Speech Presumption, Plosive or Silence* and *Possible Speech Continuation* states are introduced in order to cope with the energy variability in the observed signal and to avoid various kinds of noise. The *Speech Presumption* state avoids the automaton going into the *speech* state when the energy increase is due to an impulsive noise. However, in very noisy environment, many noises are not impulsive and with high energy. In (Karray and Martin, 2003), we propose three different energy constraints C1 considered in the automaton. We recalled here these three criteria.



Conditions	Actions
C1: Energy< *threshold energy*	A1: SiD=SiD+1
C2: Speech Duration (SD)≥64 ms	A2: SD=1
C3: Silence Duration (SiD)≥240 ms	A3: SiD=SiD+SD
	A4: SD=SD+1
	A5: SiD=1
	A6: SiD=SD=0

Fig. 2. Five state automaton.

3.1. Detection algorithm based on a SNR criterion

This algorithm was first introduced in (Mauuary and Monné, 1993). In order to decrease the energy dynamic, the log-energy is considered. The technique relies on a comparison between short-term and long-term estimates of the signal log-energy. The short-term estimate, referred to as E, is the logarithm of the mean energy of the samples computed over a window of 32 ms. The long-term log-energy estimate of the noise, referred to as LTEE, is recursively computed on the current frame n (but in the *Non-Speech* state only), by

$$\mathrm{LTEE}(n) = \mathrm{LTEE}(n-1) + (1-\lambda) \times (E(n) - \mathrm{LTEE}(n-1)) \tag{1}$$

here, λ is the forgetting factor optimized by Mauuary (1994) to 0.99. The constraint C1 is defined by the comparison of the short-term and long-term log-energy difference in dB, to an adaptive threshold referred to as the *adjusting threshold*:

$$\mathrm{C1} : E(n) - \mathrm{LTEE}(n) > \text{ adjusting threshold.} \tag{2}$$

This criterion will be referred to as the SNRC criterion. We have shown previously (Martin, 2001) that it is well adapted in high SNR environments.

3.2. Detection algorithm based on non-speech statistical criterion

First, let us assume that the non-speech log-energy distribution is a Gaussian distribution. The reader interested in the reasons behind this assumption may refer to Martin (2001). The non-speech log-energy statistics (mean and standard deviation) are estimated in the *Non-Speech* state. The mean is estimated recursively as in (1), and the standard deviation is estimated by

$$\widehat{\sigma}(n) = \widehat{\sigma}(n-1) + (1-\lambda)(|E(n) - \widehat{\mu}(n-1)| - \widehat{\sigma}(n-1)), \tag{3}$$

where $\lambda = 0.95$ is an optimized forgetting factor (Karray and Martin, 2003). The assumption that a given frame is a non-speech frame is tested by comparing the centered and the normalized log-energy of the frame: $r_{NS}(E(n)) = (E(n) - \widehat{\mu}(n))/\widehat{\sigma}(n)$ to an *adjusting threshold*. So, the constraint C1 is given by

$$\text{C1}: r_{NS}(E(n)) > \text{ adjusting threshold.} \tag{4}$$

This criterion will be referred to as the NS criterion.

3.3. Detection algorithm based on non-speech and speech statistical criterion

In order to consider this criterion, both non-speech and speech distributions are considered Gaussian as in (Karray and Monné, 1998). For the current considered frame, we have to deal with an assumption testing problem, where:

- H_0: the current frame is a non-speech frame,
- H_1: the current frame is a speech frame.

The decision rule considers the most probable assumption, according to the Bayesian approach. For each frame n, log-energy $E(n)$ is considered, and both maximum likelihoods $P(H_i/E(n))$ for each hypothesis $i = 0, 1$ are compared; i.e. the likelihood ratio given by $r_{NSS}\ (E(n)) = P(H_0/E(n))/P(H_1/E(n))$, is compared to 1: when it is less than 1, the frame n belongs to a non-speech segment, otherwise the frame n belongs to a speech segment. The constraint C1 is given by

$$\text{C1}: E(n) > \text{ adjusting threshold,} \tag{5}$$

where *adjusting threshold* $= s\alpha$: s is the solution of the equation $r_{NSS}\ (x) = 1$, where $x = E(n)$ and α is an interpolation factor in order to correct the estimation errors of statistics. Different values of α will be considered in order to obtain evaluation curves. To determine s, both assumptions are considered as equally distributed. Using Bayes formula, we have $r_{NSS}\ (x) = P(x/H_0)/P\ (x/H_1)$, where $x = E(n)$. The Gaussian distributions $P\ (x/H_i)$ are evaluated on $i = 0{,}1$ estimating means and standard deviations recursively as (1) and (3). In adverse conditions, the speech parts of the observed signal are corrupted by noise (ambient noise or transmission distortion, etc.). Thus, speech statistics represent the statistics of speech plus noise. This criterion will be referred to as the NSS criterion.

The three criteria have been evaluated with different databases in (Martin, 2001), following the evaluation procedure described in Section 2.3. The results show that the three criteria are very close in terms of recognition error rates. In the next section, the three criteria are evaluated in noisy environments with a noise reduction system.

4. Noise reduction

In a very noisy environment, a noise reduction system before the whole system presented in Fig. 1 can help to achieve better performances with the detection and recognition systems. Many speech enhancements techniques that could be used before the recognition stage have been studied, e.g. in (Karray and Martin, 2003; Wu et al., 1999). In this section, the aim is not to describe a new noise reduction system, but rather to present a comparative study of the three previous criteria in a very noisy environment, with noise reduction and without noise reduction.

The noise reduction system we used is introduced in (Noé et al., 2001), where two approaches are proposed: one time domain noise reduction and one frequency domain noise reduction. Both noise reductions give approximately the same speech recognition results. The frequency domain noise reduction is included in the MFCCs calculation, so the time domain noise reduction is more adapted to our whole system. The spectral density of the useful signal is estimated with a decision-directed as in (Ephraim and Malah, 1984):

$$\widehat{\gamma}_{\mathrm{s}}(k,f) = \beta\left|\widehat{S}(k-1,f)\right|^2 + (1-\beta) \times \max\left(|X(k,f)|^2 - \widehat{\gamma}_{\mathrm{b}}(k,f), 0\right), \tag{6}$$

where k and f are, respectively, the considered frame and frequency, X is the spectral density of the noisy signal, β is a threshold optimized to 0.98, $\widehat{\gamma}_{\mathrm{s}}$ is the spectral density estimation of the useful signal, and $\widehat{\gamma}_{\mathrm{b}}$ is the spectral density estimation of the noise.

In order to study the effects of the noise reduction stage, we used the evaluation procedure that we have just described previously, and ran a benchmark on two kinds of noisy environments (E_A and E_B), by letting the noise reduction either switched on (which is noted +NR) or off.

Environment E_A was characterized by many noises characterized by brief duration and high energy, as found in the GSM_L database with the signals of SNR less than 18 dB. On the other hand, E_B attempted to describe a typical environment with very high background noises. We simulated two versions of E_B by adding two kinds of noise with varying SNR on the GSM_L database part with SNR greater than 18 dB. The first type of noise was car noise, and the second type of noise was a babble noise, where several people talked in the background. We observed that the relative performances between the three criteria is similar for SNRs ranging from 0 to 15 dB. For the sake of clarity, only the results obtained for a SNR of 12.5 dB are given here.

4.1. Detection experiments

Fig. 3 shows the comparative study of the three criteria for detection, with and without noise reduction, when used in environment E_A. As mentioned before, the particularity of this environment is that it contains many noises with high energy. We observe that speech/non-speech discrimination is better when the noise reduction stage is switched off, and this holds for any of the three criteria. A more precise study of the non-speech detections by Karray and Martin (2003) agrees with these results, and shows that more noise events are detected after the noise reduction stage because of errors produced by the voice activity detection (VAD) module of the noise reduction stage. Our simple VAD detects the most energetic signals, that are assumed to be speech, but with the side effect of also taking energetic noises into consideration. Another result that can be observed in Fig. 3, is that the NSS criterion outperforms the other criteria, however the difference in performance between all the criteria remains small.

Figs. 4 and 5 show the performances of the three criteria on the database with, respectively, car and babble noise added. Unlike Fig. 3, we have now better performances for the three criteria when noise reduction is used. Noise reduction is more adapted for stationary noises than for noises characterized by brief duration and high energy, which is the case here. Notice that without noise reduction, the SNRC criterion results are very

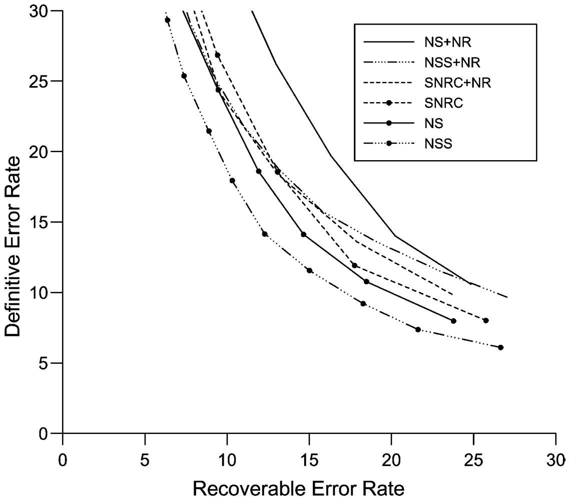


Fig. 3. Detection test: SNRC, NS, and NSS criteria, with noise reduction (+NR) and without noise reduction on the database part with SNR less than 18 dB (environment E_A).

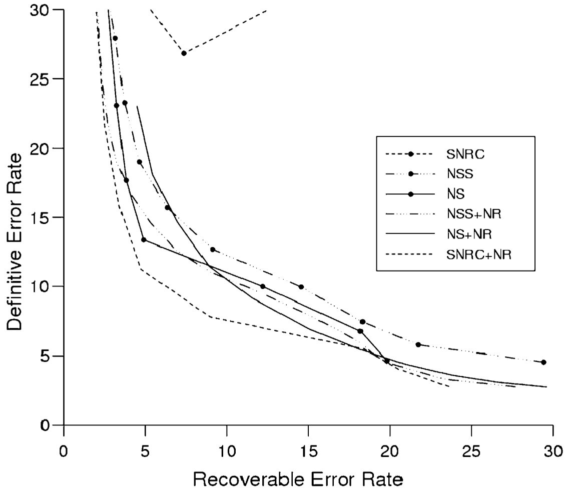


Fig. 4. Detection test: SNRC, NS, and NSS criteria, with noise reduction (+NR) and without noise reduction on the database part with a car noise added.

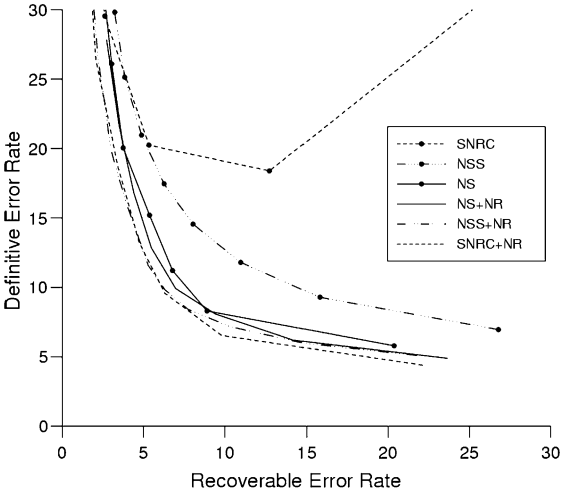


Fig. 5. Detection test: SNRC, NS and, NSS criteria, with noise reduction (+NR) and without noise reduction on the database part with a babble noise added.

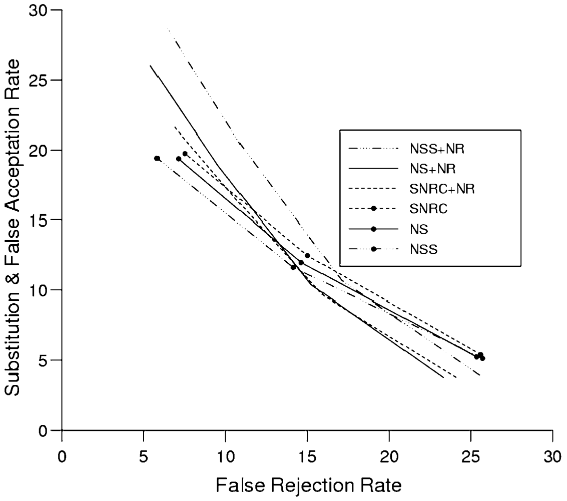


Fig. 6. Recognition test: SNRC, NS, and NSS criteria, with noise reduction (+NR) and without noise reduction on the database part with SNR less than 18 dB.

poor. On the contrary, the best results are obtained with the NS criterion, which is adequate for stationary noises, even very with high energy. With the noise reduction system turned on, the SNRC criterion outperforms the other two criteria, but the benefits remain small. Moreover, there is no difference in terms of results for the two kinds of added noises. Thus, in noisy environment, we have to take into account the noise statistics in the noise reduction system or in the speech/non-speech detection (with NS or NSS criteria).

4.2. Recognition experiments

Fig. 6 presents the recognition performances for the three criteria on the database part with SNR less than 18 dB. Fig. 3 shows that the detection results are worse with the noise reduction than without. But Fig. 6 does not show big differences between the criteria with or without noise reduction. Indeed, the detection differences come from more noise detections after the noise reduction. These detected noises are well rejected by the speech recognition system. Hence, there are no differences between the recognition results. However, with the noise reduction the speech/non-speech detection has detected more signals and so the speech recognition is more solicited. The use of the noise reduction achieves more insertions that the recognition system has to reject in order to avoid false acceptance errors. But the capability of the recognition system to reject noises is not perfect and in overall the use of the noise reduction achieves more false acceptance errors.

In order to evaluate the effect of noise reduction alone on the recognition system, in (Martin, 2001) results are presented in the case of added noises with manual segmentation instead of the detection system. In this case, the noise reduction produces better performances. Without the noise reduction, results are better with added babble noise added than with added car noise; the contrary is true when the noise reduction is turned on. Indeed, a speech-like babble noise, is more difficult to reject for the noise reduction system.

Figs. 7 and 8 present the performances of the three criteria with car and babble noise added. Notice that for both added noises, the noise reduction produces better performances. The three criteria results are similar with the noise reduction for the car and babble noise added. The improvement is statistically significant. Notice also that the SNRC criterion performances are very bad without noise reduction with car and babble noise added.

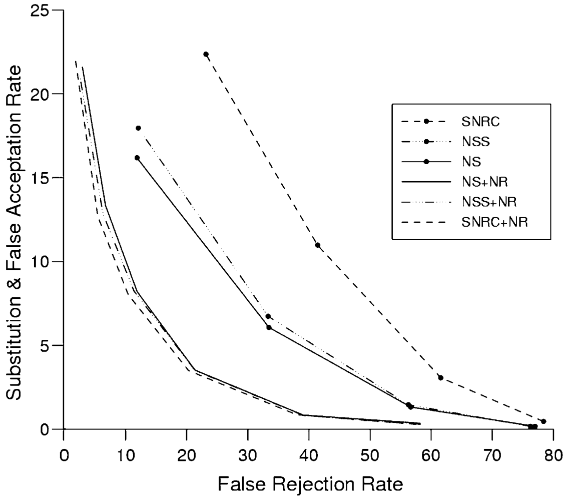


Fig. 7. Recognition test: SNRC, NS, and NSS criteria, with noise reduction (+NR) and without noise reduction on the database part with a car noise added.

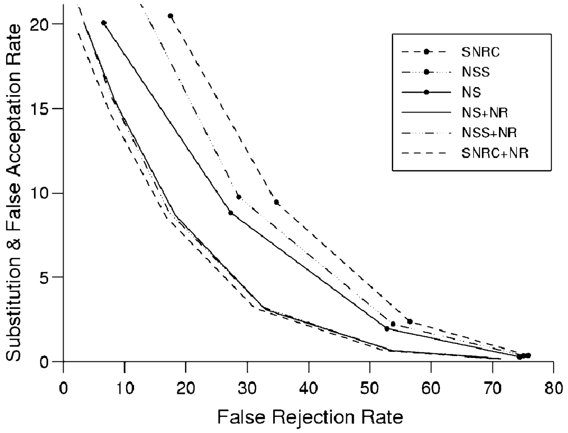


Fig. 8. Recognition test: SNRC, NS, and NSS criteria, with noise reduction (+NR) and without noise reduction on the database part with a babble noise added.

In summary, the presented noise reduction system leads to improvement for stationary noises (even with a weak SNR); and improvement is better if the noise is not babble noise. With the noise reduction the three criteria results are not significantly different. But without noise reduction, the NS criterion detects fewer noises and presents better performances. The noise reduction provides a good performance for a very noisy environment, when the noise is stationary. However, in the case of a very noisy environment with short-time energetic noises, performances stay bad. In the next section, some solutions aiming to improve the NS criterion in this case are proposed.

5. Speech/non-speech detection system using LDA applied to MFCCs

The most widely used parameter for speech/non-speech detection is the energy. This single parameter can lead to good performances, for example using the SNRC criterion (Mauuary and Monné, 1993) or the non-speech and speech energy statistics (Martin et al., 2000). But, for robustness to short-time energetic noises, most of the time the energy criterion is used with other parameters, for example pitch in (Iwano and Hirose, 1999) or entropy in (Huang and Yang, 2000). A large number of parameters can be used in the method proposed by Rabiner et al. (1977) with or without the energy. Several methods are possible to combine these parameters, like distance measurement in (Rabiner et al., 1977), classification and regression tree in (Shin et al., 2000) or data fusion methods.

In the recognition system used in this article, as in several recognition systems (Mokbel et al., 1997), the MFCCs are calculated. So the use of these coefficients does not require us to calculate more coefficients only for the speech/non-speech detection. In the case of two classes, the non-speech and the speech classes, a LDA applied to MFCCs determines a linear function to integrate all MFCCs in a single coefficient.

5.1. LDA applied to MFCC integration

The linear function *a* is calculated by LDA on both learning databases described in Section 2 using the MFCCs. This linear function is integrated into the algorithm based on the NS criterion as another condition, referred to as C4. In order to decrease the number of false detections of short-time energetic noises, C4 is added between the *Speech Presumption* and *Speech* state, described in Fig. 9.

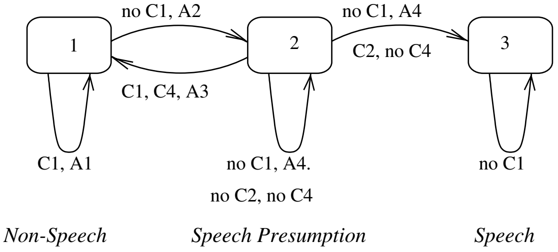


Fig. 9. Five states automaton with the new condition C4.

When the automaton is in the *Speech Presumption* state, if (i) the centered and normalized log-energy is high enough, i.e. greater than the *adjusting threshold*, (ii) speech duration is greater than 64 ms, and (iii) the MFCC linear combination obtained by LDA is less than a new threshold, referred to as the *LDA threshold*, then the automaton goes in the *Speech* state. If one of these three conditions is not realized, the automaton goes back into *Non-Speech* state (C1 and C4) or stays in *Speech Presumption* state (no C2). Thus, the new condition C4 is given by

$$\text{C4} : a.X(n) < \mathit{LDA}\ \text{threshold}, \tag{7}$$

where $X\,(n)$ is the MFCCs vector of the frame n. The*LDA threshold* is optimized on both learning databases (GSM_L and PSN_L). This new test prevents the automaton from switching to the *Speech* state, when the energy increase is due to short-time energetic noise. This new criterion will be referred to as the NS + LDA criterion.

5.2. Experiment results

The NS + LDA criterion performances compared to the performances of the NS criterion are presented following the evaluation procedure on both parts of the GSM database (different from both learning databases).

5.2.1. Detection experiments

Fig. 10 shows the NS and NS + LDA criteria performances on both GSM database parts according to SNR. The *adjusting thresholds* are noted on the curves. Notice that the NS + LDA criterion outperforms the NS criterion for both parts. The improvement is higher on the part with the SNR less than 18 dB. But on both parts, the improvement is statistically significant (according to our evaluation procedure). For one fixed *adjusting threshold* (e.g. 1.7), the NS + LDA criterion decreases the number of recoverable errors (i.e. the non-speech detections) with and without noise reduction. This decrease of recoverable errors explains the global improvement. This was expected with the new condition C4.

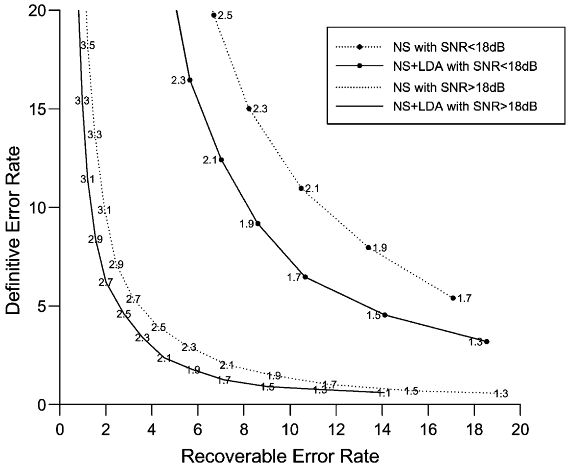


Fig. 10. Detection tests: NS and NS + LDA criteria on GSM database according to the SNR.

5.2.2. Recognition experiments

Fig. 11 presents the recognition performances of the NS and NS + LDA criteria on both parts of the GSM database. The improvement is not as high as the detection improvement. Indeed, the non-speech detections of the NS criterion are rejected by the recognition system. It is better on the part with SNR less than 18 dB, and statistically significant for a false rejection rate less than 10% (generally considered as a maximum for the user).

The aim to reduce short-time energetic noise detections is reached by the NS + LDA criterion; improvements for both speech/non-speech detection and speech recognition performances are observed. Moreover, the NS + LDA criterion that detects fewer noises reduces the whole system computational cost. Several parameters have been applied with the LDA in (Martin et al., 2001) without improvement.

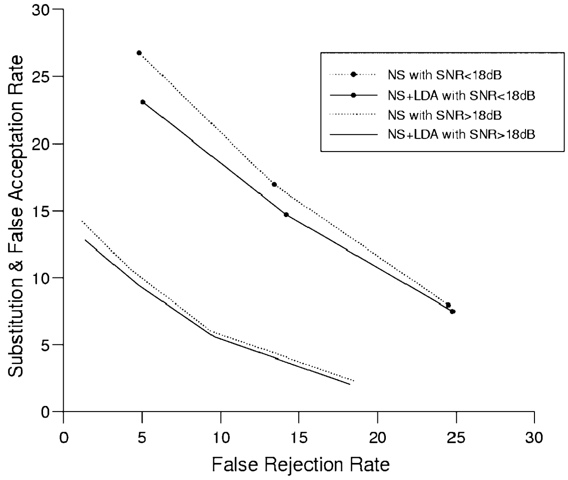


Fig. 11. Recognition test: NS and NS + LDA criteria on GSM database according to the SNR.

6. Speech/non-speech detection system using a voicing parameter

In order to discriminate the noise characterized by brief duration and high energy and speech signal, several studies use the energy with a voicing parameter. Indeed, the voiced sounds correspond to vocal cord vibrations. These vibrations can be seen as the fundamental frequency, referred to as F_0, or more generally *pitch*. In order to estimate the fundamental frequency, a zero crossing rate can be calculated and used with the energy in (Savoji, 1989; Ganapathiraju et al., 1996; Gupta et al., 1997). However, the zero crossing rates are too unstable in noisy environments (see in (Huang and Yang, 2000; Shin et al., 2000)). Thus, a precise F_0 estimation must be calculated. Many studies propose to use energy and F_0 (with or without other parameters) at all the frames as in (Kobatake et al., 1989; Junqua et al., 1991; Ramana Rao and Srichand, 1996). However the energy is a good parameter when the SNR is high enough. In order to discriminate noise with high energy and speech frames, a new combination between energy and F_0 is proposed only for energetic frames.

First we present the new voicing parameter and its integration in the speech/non-speech detection system. Next, the new criterion is evaluated following the evaluation procedure.

6.1. Voicing parameter integration

In order to obtain a precise F_0 estimation, a F_0 is calculated for voiced and unvoiced sound. The estimation method used is introduced in (Martin, 1982). The signal harmonicity is calculated by cross-correlation with a comb-function. Thus, a F_0 value is obtained every 4 ms. In order to avoid artifacts, the median is calculated, referred to as *med*. As recognition system is working on frame of 16 ms, we obtain four under-frames of 4 ms by frame. Next the difference between the current and preceding median is considered. Thus, a mean-variation estimation, referred to as $\overline{\delta med}$, is calculated for the current under-frame of 4 ms (noted m) over the N preceding under-frames:

$$\overline{\delta med}(n) = [1/N] \sum_{m=n-N}^{n} |med(m) - med(m-1)|. \tag{8}$$

This mean-variation is used as an estimation of a voicing parameter (Martin and Mauuary, 2003). This criterion is integrated into the automaton as in Fig. 12, with the new condition C4:

$$\text{C4} : \overline{\delta med}(4m) < F_0 \text{ threshold } m \in N^*. \tag{9}$$

In order to obtain a decision after every 16 ms frame, the mean-variation is considered every 4 under-frames of 4 ms (4m). Thus, this integration

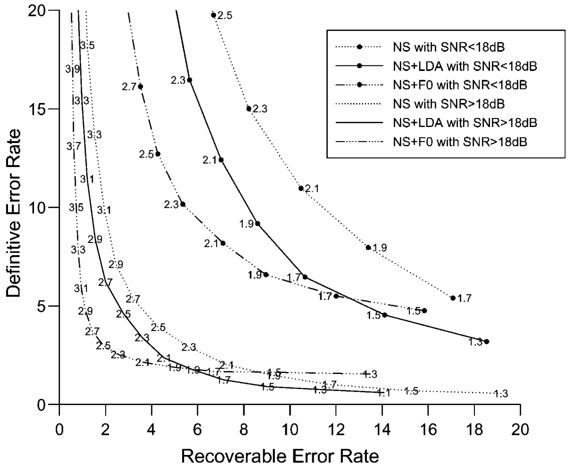


Fig. 12. Detection test: NS + F0, NS and NS + LDA criteria on GSM database according to the SNR.

avoids the automaton to go in the *Speech* state for noises characterized by brief duration and high energy, i.e. the non-speech detections must decrease. This new criterion will be referred to as the NS + F0 criterion.

6.2. Experiments

First of all, in order to test this criterion for environments with noises characterized by brief duration and high energy, the NS + F0 criterion is compared to the NS and NS + LDA criteria on both GSM database parts. As this criterion presents the best performances, it is also evaluated for stationary noise environment. So, the evaluation of the NS + F0 criterion is made with noise reduction and without noise reduction in comparison with the NS criterion on the GSM_L database part with SNR less than 18 dB, and with car and babble noise added on the GSM_L database part with SNR greater than 18 dB.

6.2.1. Detection experiments

For environment with noises characterized by brief duration and high energy, Fig. 12 shows the detection performances for the NS, NS + LDA and NS + F0 criteria on the GSM database according to the SNR. The *adjusting thresholds* are noted on the curves. The NS + F0 criterion outperforms both NS and NS + LDA criteria. The improvement is statistically significant on both database parts. For one fixed threshold (e.g. 1.9 with SNR less than 18 dB) we note a recoverable error reduction like the NS + LDA criterion, but we observe also a definitive error reduction. A precise study shows that this last reduction is due to fragmentation errors. Fragmentation errors reduction can be explained by a better estimation of the noise statistics.

Figs. 13–15 present the detection results of the NS + F0 criterion with noise reduction and without noise reduction in comparison with the NS criterion. For environment with noises characterized by brief duration and high energy, Fig. 13 shows the detection performances on the GSM_L database part with SNR less than 18 dB. We have seen that in this case the three SNRC, NS and NSS criteria performances are better without the noise reduction. However, Fig. 13 shows that the NS + F0 criterion performances are better with the noise reduction than without. So, NS + F0 criterion outperforms NS criterion with noise reduction and without noise reduction. Section 4 shows that the noise reduction creating a bigger local difference between noises with high energy and

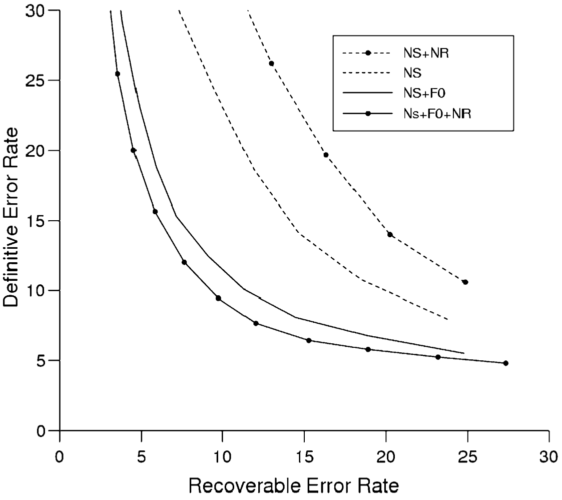


Fig. 13. Detection test: NS and NS + F0 criteria on the GSM_L database part with SNR less than 18 DB, with noise reduction (+NR) and without noise reduction.

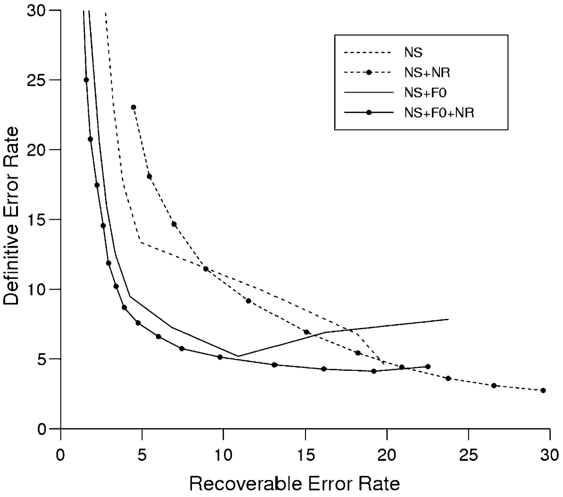


Fig. 14. Detection test: NS and NS + F0 criteria on the GSM_L database part with car noise added, with noise reduction (+NR) and without noise reduction.

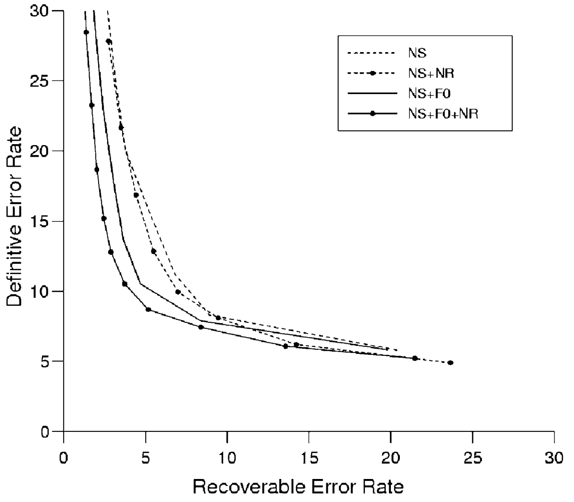


Fig. 15. Detection test: NS and NS + F0 criteria on the GSM_L database part with babble noise added, with noise reduction (+NR) and without noise reduction.

background noises allows more noise detections for the three SNRC, NS and NSS criteria. The NS + F0 criterion does not detect these noises, because of the new condition C4 that avoids these detections; this explains the improvement. For stationary noise environment, Figs. 14 and 15 show the detection performances for both NS and NS + F0 criteria on the GSM_L database part with, respectively, car and babble noise added. Both criteria performances are better with the noise reduction than without. Without noise reduction, when the thresholds are weak, a lot of noises are detected, which explains bad estimations of statistics and so the increasing of definitive errors on Fig. 14. The NS + F0 criterion outperforms the NS criterion for both added noises and with or without noise reduction. With the noise reduction turned on, the improvement compared to NS criterion is still statistically significant.

6.2.2. Recognition experiments

For environment with noises characterized by brief duration and high energy, Fig. 16 shows the recognition performances for the three NS, NS + LDA and NS + F0 criteria on the GSM database according to the SNR. Notice that the improvement of the NS + F0 criterion compared to the NS criterion is the same as the one witnessed for the NS + LDA criterion: NS + F0 and NS + LDA criteria results are similar in environment with noises characterized by brief duration and high energy. But, Fig. 12 shows that NS + F0 detects fewer noises than NS + LDA criterion, decreasing the computational cost of the whole system.

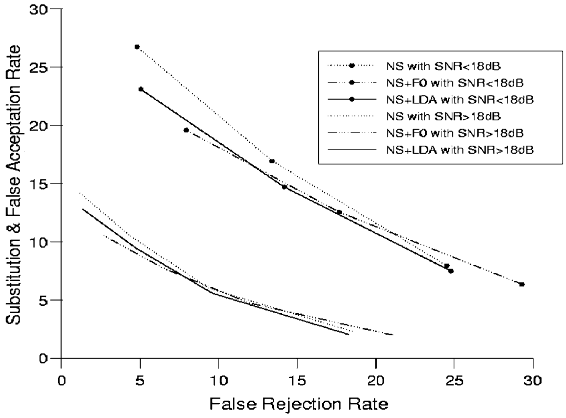


Fig. 16. Recognition test: NS + F0, NS, and NS + LDA criteria on GSM database according to the SNR.

Figs. 17–19 present recognition evaluations of both NS and NS + F0 criteria, with noise reduction and without noise reduction. Fig. 17 shows the performances of both criteria on the GSM_L

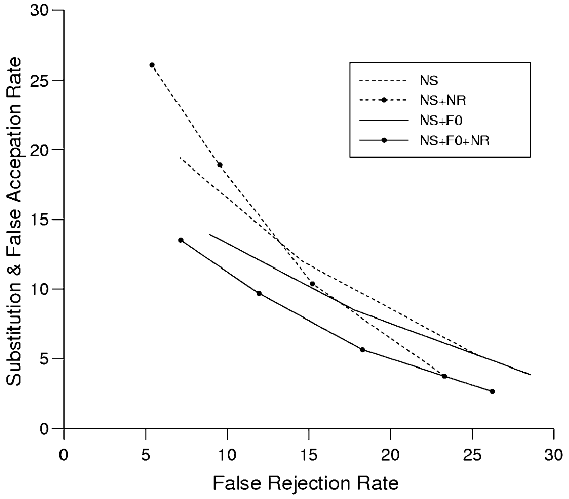


Fig. 17. Recognition test: NS and NS + F0 criteria on the GSM_L database part with SNR less than 18 dB, with noise reduction (+NR) and without noise reduction.

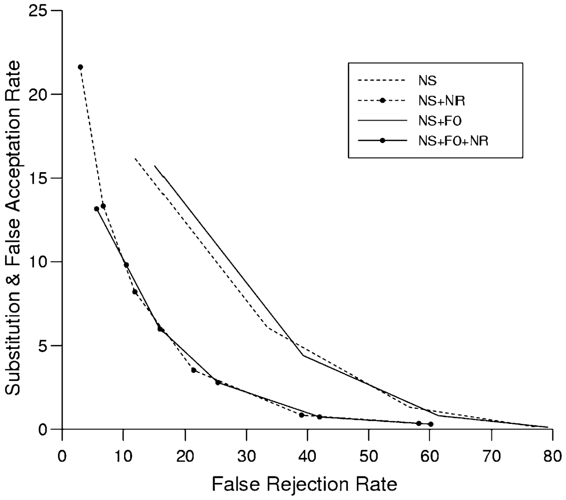


Fig. 18. Recognition test: NS and NS + F0 criteria on the GSM_L database part with car noise, with noise reduction (+NR) and without noise reduction.

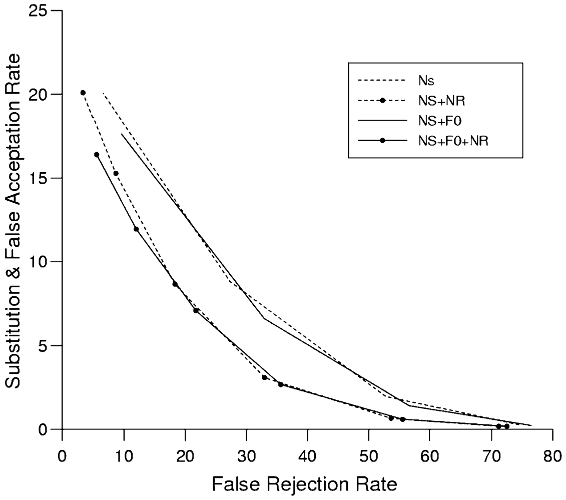


Fig. 19. Recognition test: NS and NS + F0 criteria on the GSM_L database part with babble noise, with noise reduction (+NR) and without noise reduction.

database part with SNR less than 18 dB. The NS + F0 criterion results are better with noise reduction than without. Moreover, the NS + F0 criterion outperforms the NS criterion for low false rejection rate. The improvement is statistically significant. For stationary noise environment, Figs. 18 and 19 present the recognition performances for both NS and NS + F0 criteria on the GSM_L database part with, respectively, car and babble noise added. The results with the noise reduction are better than without for both criteria and both added noises. However the NS and NS + F0 criteria results are not significantly different with noise reduction and without noise reduction, unlike the detection results. Indeed, one more time, the rejection model of the recognition system explains this. The recognition system rejects the noise detections of the NS criterion that the NS + F0 criterion does not detect. So, the NS + F0 criterion decreases the whole system computational cost.

The NS + F0 criterion outperforms the other criteria in terms of recognition results and especially in terms of detection results. With the noise reduction, the NS + F0 criterion provides good performances for stationary noises and is the best for non-stationary noises. Moreover, the NS + F0 criterion that detects fewer noises than the other criteria provides the best computational cost.

7. Conclusion

In this paper, several solutions are proposed to improve the speech/non-speech detection robustness to noise (stationary and non-stationary), in order to obtain improvements of the whole recognition system.

First, a noise reduction system is considered. The noise reduction system used improves speech/non-speech detection and speech recognition systems when the noise is stationary. However, the noise reduction system allows more detected noises when calls contain some noises characterized by brief duration and high energy. The evaluation of the three previous criteria shows that the NS criterion provides the best performances.

In order to reduce the detections of noises characterized by brief duration and high energy, two approaches are introduced. First, new speech/non-speech detection based on the energy and MFCCs is described. The MFCCs are fusioned using a linear function calculated by LDA. In environments with noises characterized by brief duration and high energy, this approach yields significant improvements for detection results. So, the whole system computational cost is reduced.

Finally, the second approach based on the energy and voicing parameter is presented. This new criterion provides the best performances. It outperforms the LDA based criterion and the three previous criteria. In environments with noises characterized by brief duration and high energy, the improvement is statistically significant for both detection and recognition results. The use of the noise reduction system with this criterion allows also the best performances. For non-stationary noises, the noise reduction with this criterion provides improvement, unlike the three previous criteria. For stationary noise, the computational cost is reduced, and the improvement given by the noise reduction is statistically significant.

In conclusion, the different proposed solutions improve the speech/non-speech detection performances according to the kind and level of noises. The use of a noise reduction system must be preferred for stationary background noise, even in very noisy environments. For environments with noises characterized by brief duration and high energy, a noise reduction system must be used according to the speech/non-speech detection system. The new introduced conditions based on MFCC or, a voicing parameter, reduce the noise detections, and so the whole system computational cost decreases. The voicing parameter-based approach can be used with noise reduction, and provides the best performances in very noisy environments (with stationary and non-stationary noises).

References

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32 (6), 1109–1121.

Ganapathiraju, A., Webster, L., Trimble, J., Bush, K., Korman, P., 1996. Comparison of energy-based endpoint detection for speech signal processing. In: IEEE Southeastcon, USA, April 1996, pp. 500–503.

Gupta, P., Jangi, S., Lamkin, A.B., Kepley, W.R., Moris, A., 1997. Voice activity detector for speech signals in variable background noise, U.S. Patent 5 649 055, July 1997.

Huang, L.-S., Yang, C.-H., 2000. A novel approach to robust speech endpoint detection in car environments. In: Internat. Conf. on Acoustics, Speech, and Signal Processing, May 2000, Turkey, pp. 1751–1754.

Iwano, K., Hirose, K., 1999. Prosodic word boundary detection using statistical modeling of moraic fundamental frequency contours and its use for continuous speech recognition. In: Internat. Conf. on Acoustics, Speech, and Signal Processing, May 1999, USA, Vol. 1, pp. 133–136.

Junqua, J.-C., Reaves, B., Mak, B., 1991. A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognizer. In: European Conf. on Speech Communication and Technology, Italia, September 1991, Vol. 3, pp. 1371–1374.

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., Martin, A., 2003. Towards improving speech detection robustness for speech recognition in adverse conditions. Speech Comm. (40), 261–276.

Karray, L., Monné, J., 1998. Robust speech/non-speech detection in adverse conditions based on speech statistics. In: Internat. Conf. on Spoken Language Processing, December 1998, Australia, pp. 1471–1474.

Kobatake, H., Tawa, K., Ishida, A., 1989. Speech/non-speech discrimination for speech recognition system under real life noise environments. In: Internat. Conf. on Acoustics, Speech, and Signal Processing, United-Kingdom, May 1989, Vol. 1, pp. 365–368.

Martin, Ph., 1982. Comparison of pitch detection by cepstrum and spectral combination analysis. In: Internat. Conf. on Acoustics. Speech, and Signal Processing, pp. 180–183.

Martin, A., 2001. Robust speech/non-speech detection for speech recognition in noisy environments, PhD, University of Rennes (in French).

Martin, A., Mauuary, L., 2003. Voicing parameter and energy based speech/non-speech detection for speech recognition in adverse conditions. In: European Conf. on Speech Comm. and Technology, September 2003, Switzerland.

Martin, A., Karray, L., Gilloire, A., 2000. High order statistics for robust speech/non-speech detection. In: European Signal Processing Conf., September 2000, Finland, pp. 469–472.

Martin, A., Damnati, G., Mauuary, L., 2001. Robust speech/non-speech detection using LDA applied to MFCC for continuous speech recognition. In: European Conf. on Speech Comm. and Technology, September 2001, Danemark, Vol. 2, pp. 885–888.

Mauuary, L., 1994. Improving the performances of interactive voice response services, PhD, University of Rennes (in French).

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

Mokbel, C., Mauuary, L., Karray, L., Jouvet, D., Monné, J., Simonin, J., Barrtkova, K., 1997. Towards improving ASR robustness for PSN and GSM telephone applications. Speech Comm. (May), 141–159.

Noé, B., Sienel, J., Jouvet, D., Mauuary, L., Boves, L., De Veth, J., De Wet, F., 2001. Noise reduction for noise robust feature extraction for distributed speech recognition. In: European Conf. on Speech Comm. and Technology, September 2001, Denmark, Vol. 1, pp. 433–436.

Rabiner, L.R., Schmidt, C.E., Atal, B.S., 1977. Evaluation of a statistical approach to voiced-unvoiced-silence analysis for telephone-quality speech. The Bell System Tech. J. 56 (March), 455–482.

Ramana Rao, G.V., Srichand, J., 1996. Word boundary detection using pitch variations. In: Internat. Conf. on Spoken Language Processing, USA, October 1996, Vol. 2, pp. 813–816.

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

Shin, W.-H., Lee, B.-S., Lee, Y.-K., Lee, J.-S., 2000. speech/non-speech classification using multiple features for robust endpoint detection. In: Internat. Conf. on Acoustics, Speech, and Signal Processing, May 2000, Turkey, Vol. 3, pp. 1399–1402.

Wu, D., Tanaka, M., Chen, R., Olorenshaw, L., Amador, M., Menendez-Pidal, X., 1999. A robust speech detection algorithm for speech activated hands-free applications. In: Internat. Conf. on Acoustics, Speech, and Signal Processing, March 1999, USA, Vol. 4, pp. 2407–2410.