ROBUST SPEECH/NON-SPEECH DETECTION IN ADVERSE CONDITIONS BASED ON NOISE AND SPEECH STATISTICS

Lamia Karray and Jean Monné

FT.CNET/DIH/DIPS
2, Av. P. Marzin, 22307 Lannion Cedex, France
E-mail : lamia.karray@cnet.francetelecom.fr

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 is a very important source of this degradation. Therefore, speech detector robustness to noise is a challenging problem to be examined, in order to improve recognition performance for the very noisy communications. Speech collected in GSM environment gives an example of such very noisy speech to be recognized. Several studies were conducted aiming to improve the robustness of speech/non-speech detection used for speech recognition in adverse conditions.

This paper introduces a robust word boundary detection algorithm reliable in the very noisy cellular network environment. The algorithm is based on the statistics of noise and speech in the observed signal. In order to decide on the binary hypotheses of noise only versus speech plus noise, we use a likelihood ratio criterion.

1. INTRODUCTION

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 were conducted in this direction [1,2,3].

High performance speech recognition requires efficient speech detection, especially in noisy environments. It is well known, indeed, that a major cause of errors in automatic speech recognition (ASR) is the inaccurate detection of the endpoints.
Many speech/non-speech detection techniques are based on energy levels. However, in real environments, the speech signal is corrupted by additive noise and this parameter may be insufficient for the correct detection of speech if the signal to noise ratio (SNR) is low.

Therefore, we developed, in previous work, a detection algorithm based on noise statistics estimation. This technique was shown to be efficient in adverse conditions [4]. For more improvement, the present work aims to enhance this statistical approach by introducing a detection algorithm based on both noise and speech statistics. The paper is organized as follows:

In section 2, we describe the detection module and two previous algorithms: one based on Speech to Noise Ratio (SNR) estimation and the other on noise statistics estimation.

Then, in section 3, we introduce the new detection algorithm based on noise and speech statistics. We also define the likelihood ratio criterion used for decision.

The detection algorithm is evaluated on a GSM mobile network database described in section 4. The adopted evaluation procedure and the obtained results are also given in this section.

Since the considered GSM database contains calls from several environments (indoor, outdoor, stopped car or running car), we summarize in section 5 the behavior of this speech/non-speech detection in each environment.

Finally, we check, in section 6, the consistence of the proposed algorithm in the case of speech over fixed networks (namely the Public Switched Network, PSN). A PSN field database is briefly described and used for this purpose.

2. SPEECH/NON-SPEECH DETECTION MODULE

The considered Speech/Non-speech Detection (SND) module consists of an adaptive five state automaton [3]. The five states are: *silence, speech presumption, speech, plosive or silence* and *possible speech continuation*. The transition from a given state to an other one is controlled by a SND algorithm and some duration constraints. These transitions between the different states determine the segment boundaries.

Two algorithms were used: one based on a SNR criterion and the other on noise statistics evaluation. These algorithms are developed in previous works and recalled in the following:

2.1. SNR Based Algorithm

For adaptive detection, the energy requirements are based on an estimation of the signal to noise ratio of the observed speech signal. The technique relies on the comparison between short-term and long-term estimates of the signal energy. This algorithm is detailed in [3,5]

2.2. Statistical Criterion

In this case, the transitions between the 5 states of the automaton are based on noise statistics estimation and duration constraints [5].

The idea consists in testing the hypothesis of noise, for each observed frame. For this purpose, we consider a normal distribution for noise energy. The noise statistics are estimated recursively, when the automaton is in the *silence* state.

By taking the noise variability into account, this statistical criterion improves the detection algorithm robustness in noisy conditions.

Since, in noisy environment the variability of speech is also very disturbing, we expand the statistical approach to the estimation of noise and speech statistics.

3. DETECTION BASED ON NOISE AND SPEECH STATISTICS

In this case, 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 distorsion, etc). Hence, the speech statistics denotes actually the statistics of *speech plus noise*.

Since the aim of speech/non-speech detection is to distinguish between noise (or non-speech) and speech frames, we consider two distributions: one for noise and one for speech. Then, we decide to which distribution belongs each frame of the observed signal.

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 consider 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:

$$r(x) = \frac{\Pr(x / H_0)}{\Pr(x / H_1)}$$

Hence, we end up with a likelihood ratio criterion.

4. EVALUATION

Before giving the results obtained using the proposed algorithm, we will first describe the evaluation conditions: recognition system, speech database and the evaluation procedure.

4.1. Experimental Data

The CNET HMM based speech recognition system, PHIL90 [6], has been used in speaker-independent context. The feature vectors considered in our experiments contain 27 coefficients. First, the energy on a logarithmic scale and the first 8 MFCC (Mel Frequency Cepstrum Coefficient) coefficients are computed on 32 ms frames; with a frame shift of 16 ms. Then, first and second derivatives of these 9 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 in order to avoid precise endpoint detection of the words to recognize. A simple Gaussian probability density function with a diagonal covariance matrix is associated to each HMM state.

The global system: acoustic analysis (MFCC), speech/non-speech detection (SND) and HMM modeling, is depicted in figure 1.

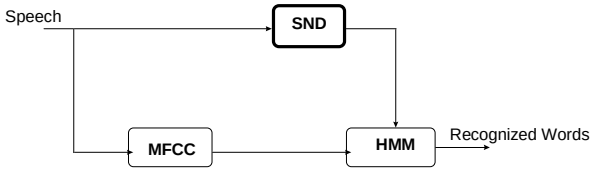


Figure 1: Global recognition system. MFCC is the acoustic analysis module and SND is the speech/non-speech detection one.

We use a laboratory GSM database of 51 words (digits and several command words) collected continuously. This means that the whole communication is recorded, including words and also silence or noise between the words.

Several call environments were considered: indoors, outdoors, stopped cars and running cars.

About 500 labeled communications are provided with almost the same proportion of each environment (26% indoors, 22% outdoors, 29% from stopped cars and 23% from running cars).

The acquisition of the whole communications results in longuer silence, so more noises. Hence, in the obtained signal, not only ambient noises are more frequent (especially in outdoor and running car calls), but also the GSM transmission effects (e.g., impulsive noises) are more disturbing. Therefore, different labels of noise and OOV (out of vocabulary) utterances are added to the initial vocabulary words. This results in a database of 35995 segments including 64% of vocabulary words, 7% of OOV words and 29% of noise (16% of ambient noises, 9% of GSM channel distorsion and 4% of remaining echoes).

The different algorithms are evaluated using this data. The evaluation procedure is described below.

4.2. Evaluation Procedure

It was shown [3] that some detection errors can be recovered by an other module (rejection module). For instance, a noise input can be rejected in the rejection module, which allows to recover the speech detector errors. 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.

4.3. Evaluation Results

Tested on the GSM database described above, this extented statistical approach results in a more robust algorithm compared to the initial one based on signal-to-noise ratio 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 table 1, we summarize the relative decrease of substitutions and false acceptance errors (which are considered as severe errors), for a given false rejection rate (about 10%).

Algorithm Performances	Substitution rate	Substitution reduction	False Acceptance rate	False Acceptance reduction
SNR	2.7	-	14.2	-
Noise Statistics	1.8	33	8.4	41
Noise & Speech Stat.	1.6	36	7.9	31

Table 1: Evaluation results in GSM environment. For a given false rejection rate (~10%), we show substitution error rates and false acceptance rates obtained by each algorithm. We also give the corresponding reductions with respect to the SNR based algorithm.

This table shows that the proposed algorithm allows noticeable improvements of the overall recognition performances.

Moreover, the overall measured decrease of error rates is actually more or less important according to how noisy is the observed signal. In the following, we will provide a detailed study of the different algorithms behavior in adverse call environments (indoors, outdoors, stopped cars and running cars).

5. RESULTS IN SEVERAL CALL ENVIRONMENT

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

The results obtained with the different speech/non-speech detection algorithms mentioned above (based on SNR, noise statistics or noise and speech estimation) are given, in figure 2, separately for each condition.

We notice that the algorithm based on noise and speech statistics and likelihood ratio criterion gives the best performance in every condition, especially when it is compared to the initial SNR based algorithm.

However, we notice different behaviors according to 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 robustness to the variations of the ambient noise characteristics (for instance, due to speed variations in the running car case).

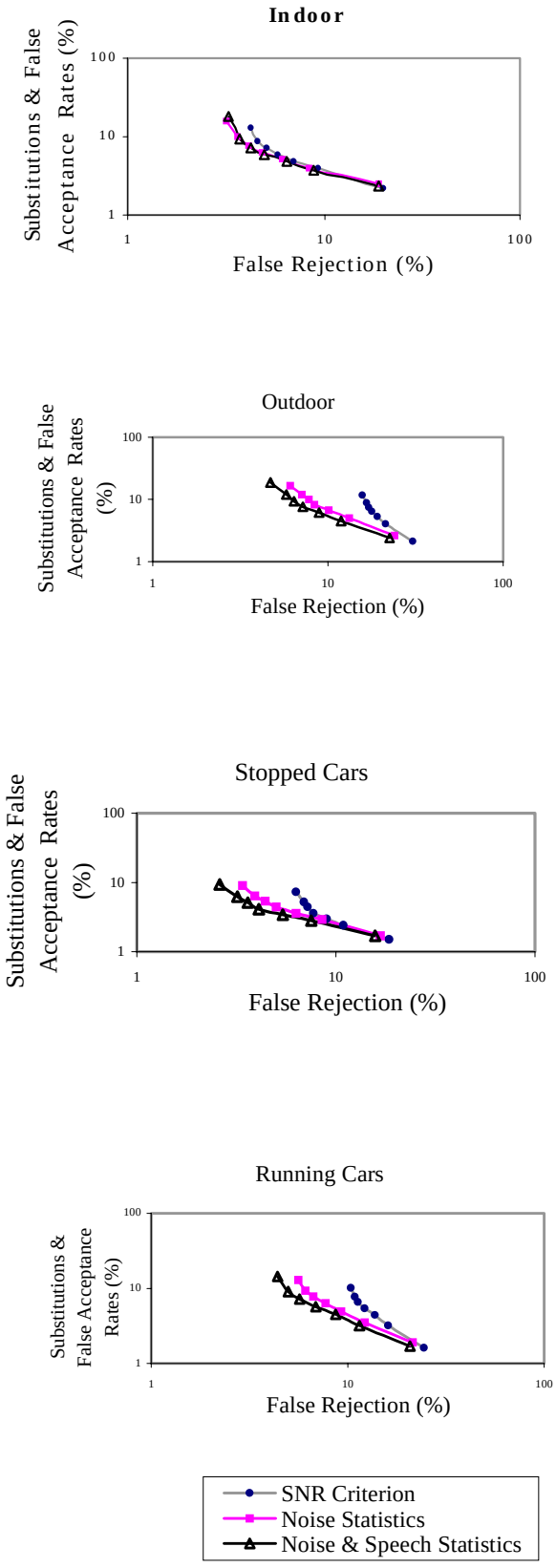
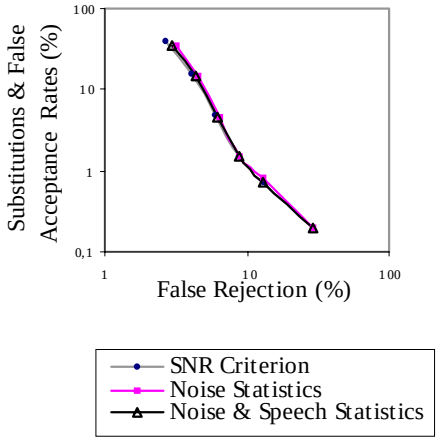


Figure 2: Evaluation of the different algorithms in adverse conditions. For each call environment, we plotted the severe error rates (False acceptance & Substitutions), function of the false rejection error rates.

6. CONSISTENCE 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. A PSN continuously recorded field database is



used [5], results are shown in figure 3.
The performances achieved with the different solutions are equivalent since the initial PSN speech to be recognized contains less ambient noise than the cellular network speech.

Figure 3: Global evaluation results in PSN environment. We plot severe error rates (false acceptance and substitutions) function of the false rejection rates (%FR).

7. CONCLUSION

In order to improve the performances of speech recognition systems, this paper deals with the speech/non-speech detection robustness to noise in wireless environment. Hence, we proposed a detection algorithm based on noise and speech (actually speech plus noise) statistics and a likelihood ratio criterion. This statistical approach takes the variations in the observed speech signal into account. Therefore, it improves the speech detection, and, consequently, the global recognizer performances. Considerable improvements are noticed, especially in very noisy call environments (outdoors and running cars).

8. REFERENCES

[1] J.C. Junqua *et al.*, ``A Robust Algorithm for Word Boundary Detection in the Presence of Noise," IEEE Trans. on SAP, Vol. 2, N. 3, pp. 406-412, July 1994.

[2] H. Agaiby and T.J. Moir, ``Knowing the Wheat from the Weeds in Noisy Speech," Proc. Eurospeech97, pp. 1119-1122, Rhodes, Greece, September 1997.

[3] L. Mauuary and J. Monné, ``Speech/non-Speech Detection for Voice Responses Systems," Proc Eurospeech'93, Berlin, pp. 1097-1100, September 1993.

[4] L. Karray, C. Mokbel and J. Monné, ``Solutions for Robust Speech/non-Speech Detection in Wireless Environment," To appear in Proc. IVTTA'98, September 1998.

[5] C. Mokbel, L. Mauuary, L. Karray, D. Jouvet, J. Monné, J. Simonin and K. Bartkova, ``Towards Improving ASR Robustness for PSN & GSM Applications," Speech Communication Journal, Vol. 23, N. 1-2, pp. 141-159, October 1997.

[6] C. Sorin, D. Jouvet, C. Gagnoulet, D. Dubois, D. Sadek, and M. Toularhoat, ``Operational and Experimental French Telecommunication Services Using CNET Speech Recognition and Text-To-Speech Synthesis," Speech Communication, Vol. 17 (3-4), pp. 273-286, 1995.