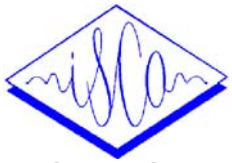

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A WAVELET DENOISING TECHNIQUE TO IMPROVE ENDPOINT DETECTION IN ADVERSE CONDITIONS

Lamia Karray and Emmanuel Polard

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

10.21437/Eurospeech.1999-521

ABSTRACT

Recognition performance decreases when automatic recognition systems are used over the telephone network, especially wireless network and noisy environments. Previous studies have shown that non efficient speech/non-speech detection is a very important source of this degradation. Hence, speech detector robustness to noise is highly required, in order to improve recognition performance for the very noisy communications.
Several studies were conducted aiming at increasing the robustness of speech/non-speech detection used for speech recognition in adverse conditions. However, despite the improvements, many segments of noise may be wrongly detected by the robust speech/non-speech detector, which increases the false acceptance errors.
Therefore, this paper introduces an efficient method to reject such false detections in order to provide a robust word boundary detection algorithm reliable in the very noisy cellular network environment. The algorithm is based on a denoising technique using a discrete wavelet transform of the detector's output segments.

Keywords: Wavelet Denoising, Speech/Non-Speech detection, Speech recognition Robustness

1. INTRODUCTION

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.

In previous works, we developed two robust detection algorithms based on noise statistics estimation [1], or on both noise and speech statistics [2]. These techniques were shown to be efficient in adverse conditions. However, we still find some wrongly detected noise segments in the detector's output.

Therefore, this paper introduces an efficient method to reject such false detections in order to provide a robust word boundary detection algorithm reliable in the very noisy cellular network environment. The algorithm is based on a denoising technique using a discrete wavelet transform of the detector's output segments. Wavelet transform was indeed shown to provide an efficient denoising technique for noisy speech signals [3,4,5,6].

The paper is organized as follows:
In section 2, we provide a brief description of the used detection system. This system is applied in the context of a GSM mobile network database.
Then, in section 3, we describe the considered denoising technique based on a discrete wavelet transform of the segments obtained by the speech/non-speech detection system.
The evaluation context and the obtained results are given in section 4.
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 enhancement in each environment.

2. SPEECH/NON-SPEECH DETECTION

The considered Speech/Non-speech Detection (SND) system consists of an adaptive five state automaton [7]. The five states are: *silence*, *speech presumption*, *speech*, *plosive or silence* and *possible speech continuation*. The transition from a given state to another one is controlled by a SND algorithm and some duration constraints. These transitions between the different states determine the segment boundaries.

Several algorithms were developed for speech/non-speech detection. They are based either on an SNR (signal to noise) criterion, or on statistical criteria (noise statistics or noise and speech statistics were used). These algorithms were developed in previous works [1,2].
In this paper, we consider the version of the algorithm based on noise statistics.

In this case, the transitions between the 5 states of the automaton are based on noise statistics estimation and duration constraints [1].
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.

This technique was applied to a noisy database collected, in adverse conditions, over the GSM network, and containing several kinds of noises. It was shown to improve the robustness of the SND considerably [1].

We noticed that the output of the speech/non-speech detection system contains mainly speech segments but also some non-speech segments (as it is shown in §4.2). Thus, we use the proposed denoising technique in order to recover some detection errors (by rejecting the wrongly detected noises).

3. WAVELET TRANSFORM DENOISING

Several previous works showed thresholding in the wavelet domain to be an effective denoising technique [3,4,5,6]. We will not provide the derivation or a detailed discussion of the wavelet transform, more details and discussions could be found in [8].
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 well localized in frequency, while fine scale wavelets are well localized in time. The advantage of this localization is that modifications can be made to the signal at particular scales without affecting, noticeably, the remainder of the signal. 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 paper, we investigate a denoising method based on a reduction or suppression of the contribution of noise while reconstructing the initial transformed segments resulting from the SND.
The idea consists in 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 among the speech segments, in the different decomposition levels.

We focus on two kinds of noise that we would like to reject: GSM network defaults (GSMN) and background noise (BN).

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 one in the second half. For example, if we consider 12 decomposition levels, M_1 is the maximum of energy in levels 1 to 6 (high frequencies), and M_2 the maximum of energy in levels 7 to 12 (lower frequencies). Figures 1 & 2 illustrate the distribution of energy in the sub-bands for a GSMN segment (figure 1) and a real speech segment (figure 2).

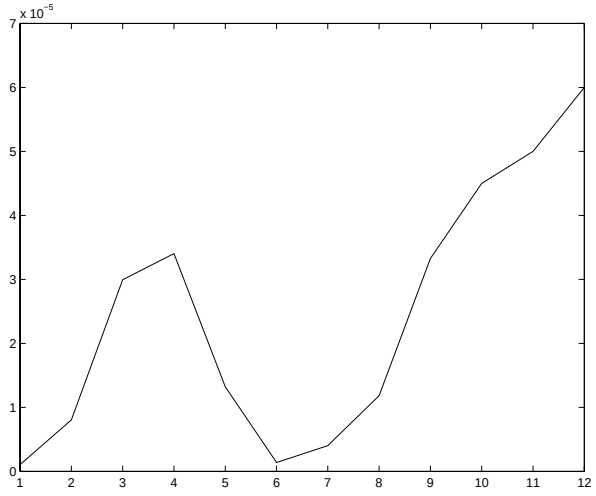

Figure 1 : Distribution of the energy in the sub-bands for a GSMN segment. M_1 is reached in level 4, and M_2 in level 12. M_1 is slightly lower than M_2.

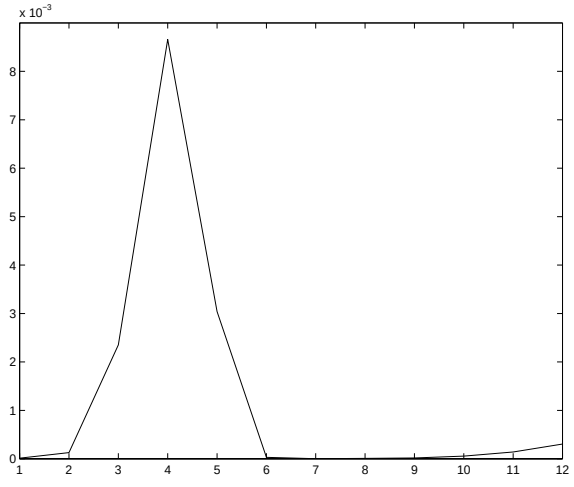

Figure 2 : Distribution of the energy in the sub-bands for a speech segment. M_1 is reached in level 4, and M_2 in level 12. M_1 is much higher than M_2.

Then, we compute the ratio: $R = \frac{M_1}{M_2}$. A statistical study of this ratio shows that:

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

This study allows 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 entirely;
2. If *2<R<5*, the segment may be corrupted by noise, we can enhance it if we reduce the contribution of the first decomposition levels (high frequency) by under-weighting them while reconstructing the segment (e.g., a weight of 0.75, instead of 1, given to levels 1 and 2);
3. If *R<2*, the segment is more likely to be noise (GSMN), we can 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.

4. EXPERIMENTAL ENVIRONMENT AND RESULTS

We apply the considered detection algorithm and the proposed denoising technique in the context of a noisy database collected, in adverse conditions, over the GSM network, and containing several kinds of noises, as it is detailed in the following.

4.1. Evaluation Context

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 longer 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, BN, 9% of GSM channel distortion, GSMN, and 4% of remaining echoes).

The SND system is applied to this data, using the 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 [9], with a decomposition depth of 12 (i.e., 12 decomposition levels).

The method is evaluated in terms of the percentage of reduction of detection errors. A complete evaluation could be performed in terms of recognition performance, in order to quantify the enhancement effect. But this is not the aim of this paper.

4.2. Evaluation Results

We apply the considered denoising technique to the segments obtained after the speech/non-speech detection using the detection algorithm mentioned above.

Despite the robustness of the speech/non-speech detection algorithm, the output of the DNS contains some remaining non-speech segments. In our example, 14.5% of the detected segments are non-speech, 25% of them are due to GSMN and 18.6% are BN.

The post-processing technique, introduced in this paper, allows a significant reduction of non-speech wrongly detected segments (particularly, GSMN and BN), as it is shown in table 1 below.

Detected Segments	**Non-speech**	**GSMN**	**BN**	**Speech**
# Rejections	1167	603	59	98
Reduction	46%	69%	9%	0.5%

Table 1: Evaluation results in GSM environment. 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 BN). However, some speech segments were also rejected. The analysis of these miss-rejected speech segments reveals that the corresponding speech is highly corrupted by noise. So, they are very unlikely to be correctly recognized. Hence, this miss-rejection would not decrease the overall recognition performances.

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

5. RESULTS FOR EACH 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.

We have shown in [2] that the considered SND system has different performances 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 speech/non-speech detection algorithm mentioned above are given, in table 2, separately for each condition. The results are given in terms of percentage of rejected segments (non-speech, GSMN, BN and speech).

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%

Table 2: Evaluation results in several call environments. We give the percentage of segments rejected by the denoising post-processing of the detector's output, with respect to the initial results of the SND system.

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.

Hence, 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 speech in quiet communications is less corrupted. So, the distinction between such noise and speech is easier.

For noisy environments, the distinction between noise and speech is more difficult. So, the application of the rejection procedure is more restricted, which explains the lower (but still important) rejection rates in such difficult conditions.

6. 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 denoising technique applied as a post-processing of speech/non-speech detection system.

The detection algorithm is based on noise statistics. The denoising technique is based on a discrete wavelet transform of the detected segments, and a localization of noise in the decomposition levels.

This method is evaluated on very noisy data collected over the wireless network (GSM).

Considerable improvements are generally noticed, since almost 50% of wrongly detected non-speech segments are rejected by the denoising process (particularly, 70% reduction of GSMN). These rejection rates are more or less important according to the kind of noise, and the call environment.

7. REFERENCES

[1] L. Karray, C. Mokbel and J. Monné, "Solutions for Robust Speech/non-Speech Detection in Wireless Environment," in Proc. IVTTA'98, pp. 166-170, September 1998.

[2] L. Karray and J. Monné, "Robust Speech/non-Speech Detection in Adverse Conditions Based on Noise and Speech Statistics," in Proc. ICSLP'98, pp. 1471-1474, December 1998.

[3] D.Donoho, "Denoising by Soft Thresholding," IEEE Trans. Information Theory, Vol. 41, pp. 613-627, 1995.

[4] S. Burley and M. Darnell, "Robust Impulsive Noise Suppression Using Adaptive Wavelet Denoising," Proc. ICASSP'97, pp. 3417-3420, April 1997.

[5] E. Burstein and W. Evans, "Wavelet Based Noise Reduction for Speech Recognition," Proc ASRU'97, pp. 111-114, December 1997.

[6] T.R. Downie and B.W. Silverman, "The Discrete Multiple Wavelet Transform and Thresholding Methods," IEEE Trans. On Signal Processing, Vol. 46, N° 9, pp. 2558-2561, September 1998.

[7] 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.

[8] M. Vetterli and J. Kovacevic, "Wavelets and Sub-Band Coding," Prentice Hall, 1995.

[9] I. Daubechies, "Orthonormal Bases of Compactly Supported Wavelets," Comm. Pure Appli. Math, vol XLI, No. 7, pp. 909-1294.