HIGH ORDER STATISTICS FOR ROBUST SPEECH/NON-SPEECH DETECTION

Arnaud Martin, Lamia Karray and André Gilloire
France Télécom R&D/DIH/DIPS,
2, Av. P. Marzin, 22307 Lannion, France
Tel: +33 2 96 05 23 10 ; fax: +33 2 96 05 35 30
e-mail: arnaud.martin@rd.frantelecom.fr

ABSTRACT

In noisy environments, a robust speech/non-speech detection is necessary for speech recognition. This paper presents a new method for speech/non-speech detection using third-order moments. The analysis of the energy third-order moment behaviour gives useful information on energy distribution. The new algorithm is compared to the one based on noise and speech statistics presented in [5]. The results show that the new algorithm outperforms the one based on noise and speech statistics only, especially in the case of noisy environments.

1 INTRODUCTION

The recognition performance decreases in very noisy environments, therefore the speech recognition system requires efficient speech/non-speech detection. Indeed, inaccurate speech/non-speech detection causes most of the errors in automatic speech recognition.

Various studies tried to increase the speech/non-speech detection performances. The principal signal-measured parameter is the energy. Most of the time, the energy is used with another parameter like the pitch [2] or the zero-crossing rate [10]. A speech/non-speech detection algorithm was developed using noise energy estimation. Then, [5] introduced another algorithm based on noise and speech energy statistics.

It is well known that the speech signal distribution is not a Gaussian distribution. So the two first order moment are not necessarily able to modelise properly the speech signal. That is why, different studies considered high order statistics are improving speech detection system. [3] proposes the use of skewness and kurtosis utilisation (i.e. the third and fourth normalised cumulants). [4] used a source separation technique in a voice activity detection. This source separation is based on the fact that the cross cumulant of two independent variables is null. [9] used the fourth order cumulant of the LPC residual in a voice activity detection system.

Since the speech energy distribution is not a Gaussian distribution, the high order statistics give a better distribution description. In this paper, we integrate a normalised third-order moment conditionally to the initial algorithm using noise and speech statistics.

This paper is organised as follows: First we will recall the previous algorithms based on noise and speech statistics. Next we will present the third order moment, and how we integrate it in the detection system. Finally we will present the evaluation of this new criterion, and the advantage of this approach.

2 PREVIOUS ALGORITHMS

The two previous speech/non-speech detection algorithms are based on an adaptive five states automaton [6]. The five states are: *silence*, *speech presumption*, *speech*, *plosive or silence* and *possible speech continuation*. The transition from one state to another is controlled by a test on the signal features. These transitions between states and some duration constraints determine the endpoints segmentation.

We recall hereafter the two previous criteria based on noise statistics and both noise and speech statistics.

2.1 Noise Statistical Criterion

The transitions between the states of the automaton are based on noise statistics estimation. We assume that the noise energy has a normal distribution. The noise energy mean and standard deviation are estimated recursively in the *silence* state of the automaton. Then we test the hypothesis of *silence* (noise) state, for each observed frame. The consideration of speech statistics improves the decision, especially in noisy environments.

2.2 Noise And Speech Statistical Criterion

We consider here both noise and speech energy means and standard deviations, to control the transitions between the states of the automaton. The noise statistics are still estimated in *silence* state, whereas speech statistics are estimated in *speech* state. This approach comes from Bayes approach. We test the two hypotheses:

H_0: the observed frame is a noise frame (or non-speech)

H_1: the observed frame is a speech frame.

We consider now two different normal distributions, one for noise and one for speech. Hence, we compare the two likelihood $P(H_i/x)$ for $i = 0$ or 1, where x is the observed frame. Assuming the two hypotheses equally distributed, and using Bayes formula, the problem is reduced to a comparison to 1 of the likelihood ratio:

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

3 THIRD-ORDER MOMENT CRITERION

3.1 High Order Statistics Estimation

The "classical" r^{th}-order moment estimation of the energy is the arithmetic estimation:

$$\hat{\mu}_r(n) = \frac{1}{n}\sum_{i=1}^{n} x_i^r,$$

where x_i is the signal energy of the i^{th} frame, and n is the number of frames. But this estimation does not take into account the non-stationarity of signal. Hence, we use the estimation on exponential windows. This is equivalent to weighted frames with time decreasing weights. Then, for an observed frame n, the r^{th}-order moment estimation is defined as:

$$\hat{\mu}_r(n) = \hat{\mu}_r(n-1) + (1-\lambda)(x_n^r - \hat{\mu}_r(n-1)),$$

where λ is the forgetting factor. The supposed level of the signal stationarity determines the factor λ, and hence the number of considered frames used to calculate the statistics. Contrary to the arithmetic estimator, this estimator has bias, we have:

$$E[\hat{\mu}_r(n)] = (1-\lambda^{n+1})\mu_r,$$

where μ_r is the theoretic r^{th}-order moment. We note that this estimator is asymptotically without bias, when λ tends towards 1. The standard deviation formula for this estimator is difficult to calculate in the general case. A study of this estimator for high n values is available in [7]. The author demonstrated that this estimator is consistant, with a decreasing speed when the moment order increases.

3.2 The Third-Order Moment

If we assume the quantity mean equals 0, the normalised third-order moment is exactly the skewness. Some numerical considerations show that third-order statistics describe well the energy distribution, and that the estimator standard deviation is low enough. Hence, we consider the normalised third-order moment, defined as:

$$\hat{m}_3(n) = \frac{\hat{\mu}_3(n)}{\hat{\sigma}^3(n)},$$

where $\hat{\mu}_3$ and $\hat{\sigma}$ are respectively, the energy third-order moment and standard deviation estimation. They are estimated like in the previous section with exponential windows.

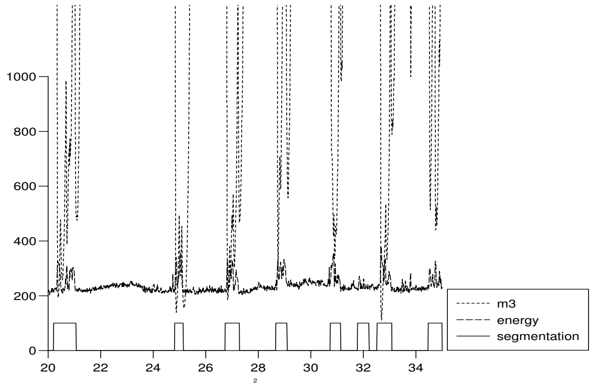

Figure 1: Third-order moment representation

We plot on Figure 1, the energy, the third-order moment and the manual speech segmentation for a very noisy signal. $\hat{m}_3$ is calculated with a factor $\lambda = 0.9$. We can see that $\hat{m}_3$ decreases considerably during the speech periods.

3.3 Third-Order Moment Integration

We integrate now this parameter in the algorithm based on noise and speech statistics. We calculate $\hat{m}_3$ with two different forgetting factors, one ($\lambda_{st} = 0.9$) that gives a short term estimation of the third-order moment, another ($\lambda_{lt} = 0.995$) that gives a long term estimation. The long term estimation is calculated recursively on the *silence* state of the automaton. We calculate the ratio of $\hat{m}_3$ obtained with $\lambda_{st} = 0.9$ and with $\lambda_{lt} = 0.995$. Then for every state, when we are in the hypothesis H_1 for the based algorithm noise and speech statistics, we compare this ratio to a threshold. The threshold is defined as follows:

$$\hat{T}(n) = \hat{T}(n-1) + (1-\lambda_T)(C.ratio(n-1) - \hat{T}(n-1)),$$

where $ratio(n)$ is the moment's ratio estimation, in a speech frame n, $C > 1$, and λ_T is a forgetting factor. Hence, $\hat{T}(n)$ overestimates the moment's ratio.

This approach assumes that both noise and speech third-order moments are different, and that noise is more stationary than speech. First test assumes that noise and speech distribution are Gaussian, and so noise and speech third-order moment are null. But this is an approximation. We consider third-order moment to improve the decision with the second test. Hence we describe the noise and speech periods better, and eliminate the wrong noise detection.

4 Evaluation

We used two databases for the evaluations. One test is performed to evaluate the segmentation obtained by this algorithm, and another is conducted to evaluate the resulting speech recognition performances. The speech recognition system used is the CNET HMM-based system [8]. We compare this algorithm with the initial

New-*Initial*	Omis.	Inser.	Reg.	Frag.
vocabulary	61-*60*	2578-*2628*	335-*348*	28-*27*
out-of-voca.	57-*42*			158-*156*

Table 1: speech segmentation errors for threshold 1.7 - PSN database

algorithm using noise and speech statistics. First we describe the used databases, next we present the segmentation test and the results. Finally we present the recognition test and the obtained results with the new algorithm. Manual segmentation gives 67% of vocabulary words, 11% of out-of-vocabulary words and 22% of noise.

4.1 Databases

A first database includes 1000 phone calls to an interactive voice response service giving movie programs. It's recorded over PSN (Public Switched Network). The obtained corpus contains 25 different words.

The second database is a laboratory GSM database of 51 words. Several call environments are considered: indoor, outdoor, stopped car and running car. Manual segmentation gives 68% of vocabulary words, 4% of out-of-vocabulary words and 28% of noise.

4.2 Segmentation Tests

The segmentation test is a comparison with a manual segmentation of the speech and noise periods. Hence we distinguish between the vocabulary words, out-of-vocabulary words and several kinds of noise. We consider different errors, the omissions (a vocabulary or out of vocabulary word not detected), the insertions (detected silence), the regrouping (several word detected as one) and the fragmentation (one word detected as several) [6]. With a view to the recognition, we class the errors, the recoverable errors representing errors that the recognition system can reject (insertion and noise and out-of-vocabulary word detection), and the definitive errors (omissions, regrouping and fragmentation). We obtain the curves for different thresholds of the initial algorithm, and we plot definitive errors function of recoverable errors.

Figure 2 presents the segmentation results on the PSN database. We note that for a given threshold the new algorithm gives more definitive errors, but less recoverable errors. For example for threshold 1.7, Table 1 shows that we have less insertion errors for the same omission errors, and more regrouping errors for the out-of-vocabulary words.

Figure 3 shows the test results on GSM database for the four environments. We notice that the new algorithm shows less recoverable and definitive errors that the initial one. The difference for the definitive errors comes from the regrouping errors. When we did the tests on the different environments, we note that the difference was more important for noisy environments (outdoor, running car).

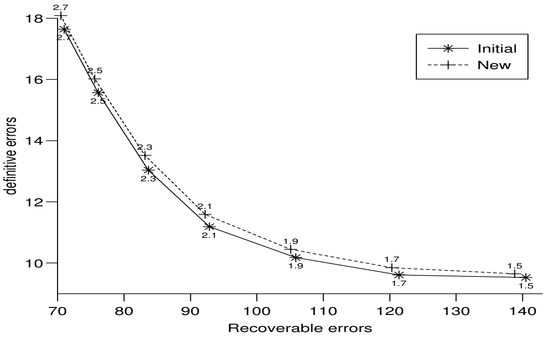

Figure 2: Segmentation Tests - PSN base

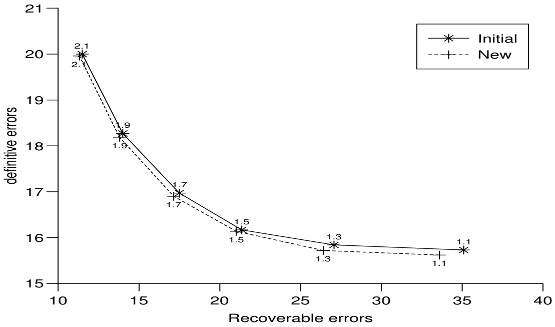

Figure 3: Segmentation Tests - GSM base

4.3 Recognition Tests

The recognition system used at CNET [8] is based on HMM. We obtain the curves with different rejection thresholds. We plot substitution and false acceptance rates (respectively, vocabulary word recognised as another vocabulary word, out-of-vocabulary word recognised as vocabulary word) function of false rejection rates (rejected vocabulary word).

The recognition test for the PSN database shows that both algorithms are equivalent. The new algorithm gives more definitive errors on this database, as we have seen in the previous section, but the difference does not change the recognition results.

On the GSM database, Figure 4 shows that the new algorithm gives a small improvement for the two considered thresholds. The improvement is minimal. The difference of errors comes from the regrouping errors, that are not important on this database. Figure 5 shows that improvement increases in noisy environments like running car.

4.4 Discussion

The obtained results show that the new algorithm is slightly more robust for noisy environments. This is due to the errors insertion that are less important. The third-order moment allows to reject more noise and

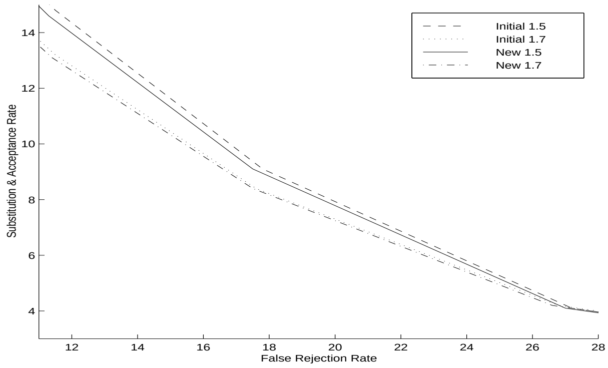

Figure 4: Recognition Tests - GSM base

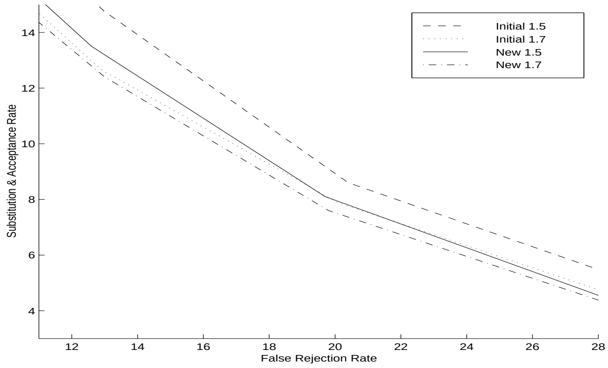

Figure 5: Recognition Tests - GSM base running car

silence without increasing omission errors. However the small improvement is not very significant on these databases. We can explain this small improvement by the fact that the initial algorithm describes correctly the noise and speech energy distribution without the high order moment. The difference between the third-order moment in noise and speech periods permits to confirm a speech period, but not to decrease omission or fragmentation errors (definitive errors). The cost of this method is not important, and this approach is interesting for very noisy environments.

5 Conclusion

We have integrated the energy third-order moment in the speech detection algorithm based on noise and speech statistics. The moment ratio allows to improve the comparison between noise and speech energy distributions. Segmentation and recognition test results show a small improvement for the used recognition system especially in noisy environments. However, this improvement is not significant. But this approach is interesting in very noisy environments, and does not decrease the performance for environment without background noise.

We integrate the energy third-order moment conditionally to the initial algorithm decision. It is possible to integrate this new criterion differently. For instance, the new criterion can be seen as another decision to decrease the omission and fragmentation errors. In a further study, we will introduce this new decision using some decision fusion method.

References

[1] F. Beritelli, S. Casale and K. Cavallaro, "A Multi-Channel Speech/Silence Detector based on Time Delay Estimation and Fuzzy Classification," ICASSP, Phoenix, Arizona, USA, March 1999, vol. 1, pp. 93-96.

[2] K. Iwano and K. Hirose, "Prosodic word boundary detection using statistical modeling of moraic fundamental frequency contours and its use for continuous speech recognition," ICASSP, Phoenix, Arizona, USA, March 1999, vol. 1, pp. 133-136.

[3] G. Jacovitti, P. Pierucci and A. Falashi, "Speech Segmentation and Classification Using Higher Order Moments," Eurospeech, Geneva, Italy, 1991, pp. 1371-1374.

[4] N. Doukas, P. Naylor and T. Stathaki, "Voice Activity Detection Using Source Separation Technique," Eurospeech. pp. 1099-1102, 1997.

[5] L. Karray and J. Monné, "Robust speech/non-speech detection in adverse conditions based on noise and speech statistics," ICSLP, Sydney, Australia, December 1998, vol. 4, pp. 1471-1474.

[6] L. Mauuary and J. Monné, "Speech/non-Speech Detection for Voice Responses Systems," Eurospeech, Berlin, Germany, September 1993, vol. 2 , pp. 1097-1100.

[7] P. McCullagh, Tensor Methods in Statistics. Charpman and Hall, 1987.

[8] C. Mokbel, L. Mauuary, L. Karray, D. Jouvet, J. Monné, J. Simonin, K. Bartkova, "Towards improving ASR robustness for PSN and GSM telephone applications," Speech communication. vol. 23, pp. 141-159, May 1997.

[9] E. Nemer, R. Gourbran and S. Mahmoud, "The fourth-order cumulant of speech signals with application to voice activity detection," Eurospeech, Budapest, Hungary, September 1999, vol. 5, pp. 2391-2394.

[10] M.H. Savoji, "a Robust algorithm for Accurate Endpointing of Speech Signals," Speech communication. vol. 8, pp. 4560, January 1989.