ROBUST SPEECH/NON-SPEECH DETECTION USING LDA APPLIED TO MFCC

Arnaud Martin, Delphine Charlet, Laurent Mauuary

France Télécom R&D
DIH/IPS, 2 av. Pierre Marzin
22307 Lannion Cedex - FRANCE
arnaud.martin@rd.francetelecom.fr

ABSTRACT

In speech recognition, a speech/non-speech detection must be robust to noise. In this work, a new method for speech/non-speech detection using a Linear Discriminant Analysis (LDA) applied to Mel Frequency Cepstrum Coefficients (MFCC) is presented. The energy is the most discriminant parameter between noise and speech. But with this single parameter, the speech/non-speech detection system detects too many noise segments. The LDA applied to MFCC and the associated test reduces the detection of noise segments. This new algorithm is compared to the one based on signal to noise ratio (SNR) [1].

1. INTRODUCTION

In a very noisy environment, the recognition performance decreases in part due to imperfect speech detection, therefore an efficient speech/non-speech detection is crucial. Indeed, in a noisy environment, the speech/non-speech detection system detects too many noises, and causes errors in automatic speech recognition [2].

The most widely used parameter for speech detection is the energy. This single parameter achieves good performance, for example using the SNR [1] or the noise and speech energy statistics [2]. For robustness to noise, most of the time the energy is used with other parameters, for example pitch [3] or entropy [4]. A large number of parameters can be used [5] with or without the energy. Several methods are possible to combine this large number of parameters like distance measure [5] or data fusion methods like classification and regression tree [6].

In several recognition systems [7], MFCC are calculated. So using these coefficients do not require to calculate more coefficients. In the case of two classes, the noise and the speech classes, a LDA applied to MFCC determines a linear function to integrate all MFCC like a single coefficient.

In this work, we use the LDA applied to MFCC, in order to improve the energy-based speech/non-speech detection. This paper is organised as follows: section 2 recalls the previous algorithm based on SNR. Section 3 reviews the LDA, and presents its integration in the speech detection system. Finally, section 4 presents the evaluation of this new algorithm.

2. ALGORITHM BASED ON SNR

In previous work [1], the speech/non-speech detection algorithm is based on an adaptive five state automaton. 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 the frame energy and some duration constraints, see figure 1. Estimations for a long-term and a short term signal energy are compared to an adaptive threshold, referred as *threshold energy*. This test and the duration constraints determine the endpoint of the detection.

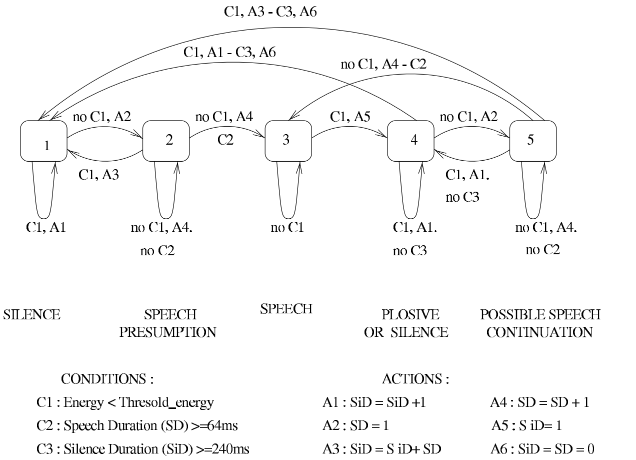

Fig. 1. Five state automaton

The three states: *speech presumption*, *plosive or silence* and *possible speech continuation* are introduced in order to cope with the energy variability in the observed speech (within-word silence) and to avoid various kinds of noise.

Hence, the *speech presumption* state avoids the automaton to go in the *speech* state when the energy increase is due to an impulsive noise. But when the energy is high and the automaton is in this state during more that 64ms, it goes in the *speech* state.

The noise & speech statistics-based algorithm [2], noted N&S STAT, works on the same automaton, but the transitions are controled with the noise and speech energy estimation.

3. LDA CRITERION

3.1. LDA

This method discriminates classes. Here, there are only two classes, the noise class and the speech class. The principle is to find a linear function $\mathbf{a}$ maximizing between-class variance and minimizing within-class variance.

The between-class covariance matrix is noted $\mathbf{E}$, the within-class covariance matrix $\mathbf{D}$ and the global covariance matrix $\mathbf{T}$. The Huyghens decomposition formula gives:

$$\mathbf{a}^*\mathbf{T}\mathbf{a} = \mathbf{a}^*\mathbf{D}\mathbf{a} + \mathbf{a}^*\mathbf{E}\mathbf{a}.$$

So, the linear function $\mathbf{a}$ is such as $\mathbf{a}^*\mathbf{D}\mathbf{a}$ is minimal and $\mathbf{a}^*\mathbf{E}\mathbf{a}$ is maximal, i.e.:

$$f(a) = \frac{\mathbf{a}^*\mathbf{E}\mathbf{a}}{\mathbf{a}^*\mathbf{T}\mathbf{a}}$$

is maximal. We have to solve:

$$\mathbf{T}^{-1}\mathbf{E}\mathbf{a} = \lambda\mathbf{a}, \tag{1}$$

with $\mathbf{a}^*\mathbf{T}\mathbf{a} = 1$. As there are only two classes, $\mathbf{E}$ is such as:

$$\mathbf{E} = \mathbf{c}\mathbf{c}^*,$$

with

$$c_j = \frac{\sqrt{n_1 n_2}}{n_1 + n_2}(\bar{x}_{1j} - \bar{x}_{2j}),$$

where n_1 is the number of noise frames, n_2 the number of speech frames, $\bar{x}_{1j}$ is noise j^{th} MFCC mean, and $\bar{x}_{2j}$ is speech j^{th} MFCC mean. Hence the equation (1) is:

$$\mathbf{T}^{-1}\mathbf{c}\mathbf{c}^*\mathbf{a} = \lambda\mathbf{a},$$

and $\mathbf{a} = \mathbf{T}^{-1}\mathbf{c}$ is the only linear function.

3.2. LDA applied to MFCC integration

The linear function $\mathbf{a}$ is calculated on two learning databases described in section 4.1. We integrate this linear function, obtained by LDA applied to the MFCC, in the algorithm based on SNR. We want to decrease the number of false detection of noise. Hence we add another condition between the *speech presumption* and *speech* state: C4, see figure 2. When automaton is in *speech presumption* state, if the SNR is high enough, i.e. energy is superior to *threshold energy*, speech duration is superior to 64ms, and the MFCC linear combination, obtained by LDA, is inferior to a new threshold, referred as *threshold LDA*, the automaton goes in the *speech* state. If one of this three conditions is not realized, the automaton goes back in *silence* state (C1 and C4), or stay in *speech presumption* state (C2).

This new test avoids the automaton to go in the *speech* state, when the energy increase is due to noise.

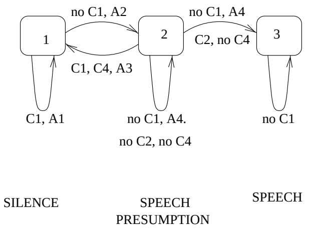

Fig. 2. Five states automaton with the new condition C4.

4. EXPERIMENTS

Two learning databases are used to calculate the linear function by LDA applied to MFCC, and two other databases are used for evaluation. Both classes, noise class and speech class, are determined by the manual segmentation of the two first databases. The noise segments constitute the noise class, and the vocabulary words and the out-of-vocabulary word segments constitute the speech class. To evaluate the detection system, LDA-based algorithm performance is compared to those of SNR-based and N&S STAT-based algorithm performance. Evaluation is made, first in term of detection errors, and then in term of recognition errors. The speech recognition system used, is an HMM-based system [7].

First the databases are described. Next the detection experiments and the recognition experiments are presented.

4.1. Databases

Two learning databases are used to calculate the linear function by LDA applied to MFCC. The first database includes 1000 phone calls to an interactive voice response service giving movie programs. It was recorded over PSN (Public Switched Network). The corpus contains 25 different words. The second database is a laboratory GSM database consisting of 51 words, including 390 phone calls. Manual segmentation on the learning databases gives 63% of vocabulary word segments, 9% of out-of-vocabulary word segments and 28% of noise segments.

Two other laboratory databases and are used to evaluate the speech/non-speech detection system.

One of these databases, referred to as PSN database, recorded over PSN, contains 68 words based on 180 phone calls and contains 68 words. This base is divided into two parts: in one part speakers repeat the words, referred to as repeat part, in the other part speakers read the words, referred to as read part. Manual segmentation gives 91% of vocabulary word segments, 3% of out-of-vocabulary word segments and 6% of noise segments.

The second database, referred to as GSM database, was recorded over GSM and contains 65 words. The 390 phone calls came from different environments: indoor, outdoor, stopped car and running car. Outdoor and running car calls are more noisy than indoor and stopped car calls, but this repartition does not correspond exactly to noise level. This database is divided into two parts: noisy phone calls (SNR is inferior to 15dB), and quiet phone calls (SNR is superior to 15dB). Manual segmentation gives 85% of vocabulary word segments, 3% of out-of-vocabulary word segments and 11% of noise segments.

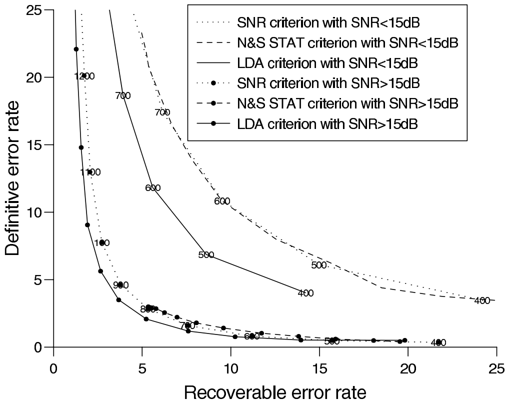

Fig. 3. Detection Tests on GSM database.

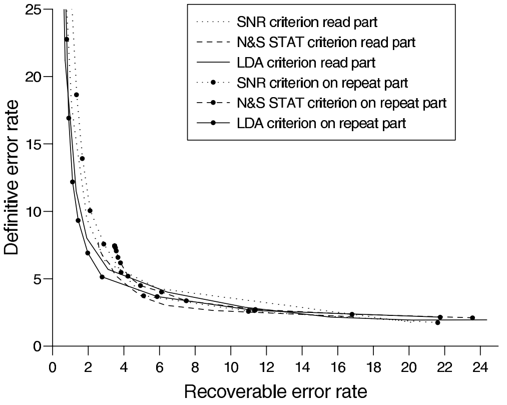

Fig. 4. Detection Tests on PSN database.

4.2. Detection Experiments

To evaluate the detection system, first we compare automatic speech segment detection to manual segmentation of speech and noise periods. Hence we distinguish between the vocabulary words, out-of-vocabulary words and several kinds of noise. Different errors are considered:
- omission: a vocabulary word or an out-of-vocabulary word is not detected,
- insertion: a noise (or silence) segment is detected, as speech
- regrouping: several words are detected as one,
- fragmentation: one word is detected as several.

Noise or out-of-vocabulary detections can be rejected by the recognition system. These errors are called *recoverable error*. The omission, regrouping and fragmentation errors unavoidably produce recognition errors. These errors are called *definitive error*. The recoverable and definitive error rates are calculated with respect to the total number of speech segment (vocabulary and out-of-vocabulary words manual segments).

To compare both criteria, SNR-based, N&S STAT-based and LDA-based, definitive errors according to recoverable errors are plotted for different energy thresholds, *threshold-energy*. The *threshold LDA* is optimized on the learning databases.

Figure 3 presents detection results on the GSM database for both parts, SNR inferior and superior to 15dB. We observe that the LDA criterion outperforms the SNR criterion and N&S STAT criterion. The difference between the LDA criterion and both other criteria is more important on the noisy part (inferior to 15 dB). But on both parts, the improvement is statistically significant.

Figure 4 shows detection results on the PSN database, for both parts, read and repeat parts. On the read part, improvement is statistically significant. On the repeat part, improvement is small but not statistically significant.

4.3. Recognition Experiments

Recognition experiments were conducted using an HMM-based speech recognition system [7]. Curves are obtained by varying the rejection threshold, for the *threshold energy* giving the minimum recognition errors. Substitution errors (vocabulary word recognized as another vocabulary word), false acceptance errors (noise or out-of-vocabulary word recognized as vocabulary word) and false rejection errors (rejected vocabulary word) are considered. To compare the three criteria, substitution and false acceptance error rate according to false rejection error rate is represented. False rejection error rate is calculated with respect to the vocabulary word manual segments, and substitution and false acceptance error rate with respect to the total number of manual segments.

Figure 5 shows recognition results on the GSM database for both parts, SNR inferior and superior to 15dB. On this database false rejection error rate is high. LDA criterion outperforms SNR and N&S STAT criteria, especially in noisy part. For the noisy part (SNR inferior to 15dB), the improvement is statistically significant. However the improvement is not statistically significant for the quiet part (SNR superior to 15dB), for a false rejection error rate between 6 and 15%.

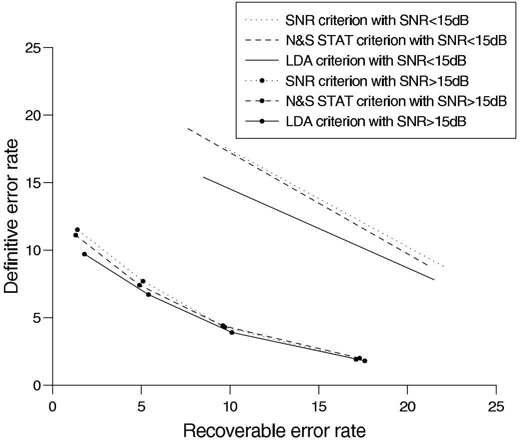

Fig. 5. Recognition Tests on GSM base

The results on the PSN database, Figure 6, show that the curves for SNR, N&S STAT and LDA criterion are crossing. The difference between the LDA criterion and both other criteria is not significant. But the LDA criterion does not decrease the performance in this environment, for a false rejection error rate between 6 and 15%. Detection experiments show that there are no improvement in this case, so no improvements are expected for recognition experiments. However the detection borders are not evaluated in this study. The precision of the detection borders can improve the recognition performance.

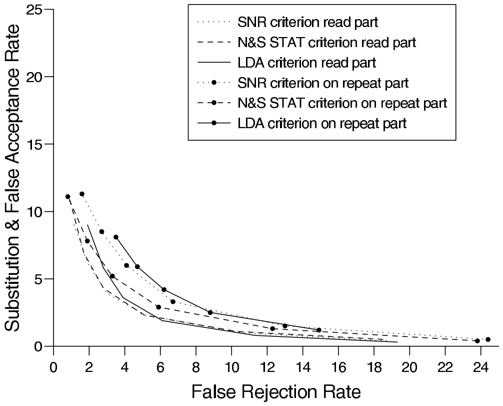

Fig. 6. Recognition Tests on PSN base

5. CONCLUSION

In this work energy and MFCC are used for Speech/Non-Speech detection. The MFFC are fusioned using a linear function calculated by Linear Discriminant Analysis. The integration of the LDA applied to MFCC in the detection system gives significant improvements for detection results especially in noisy environments. For recognition results, the improvement is significant only in noisy environments. The LDA criterion decreases the detection of noise segments. When combined with a good rejection at the recognition system level, to decrease the noise detections is unnecessary to improve recognition performance. In the noisy environment, the rejection is not efficient, so to decrease the noise detections improve recognition performance. However the LDA criterion computational cost is inferior to rejection computational cost at the recognition system level. This new criterion outperforms the SNR and N&S STAT criteria, and gives significant improvement.

We used here a linear function calculated by LDA applied to MFCC and integrated it with a new test, we are investigating other combinations of this test.

6. REFERENCES

[1] L. Mauuary and J. Monné, “Speech/non-speech Detection for Voice Response Systems,” in *Eurospeech’93*, Berlin, Germany, 1993, pp. 1097–1100.

[2] A. Martin, L. Karrray, and A. Gilloire, “High Order Statistics for Robust Speech/Non-Speech Detection,” in *Eusipco*, Tampere, Finland, Sept. 2000, pp. 469–472.

[3] K. Iwano and K. Hirose, “Prosodic Word Boundary Detection Using Statistical Modeling of Moraic Fundamental Frequency Contours and its Use for Continuous Speech Recognition,” in *ICASSP’99*, Phoenix, USA, May 1999, vol. 1, pp. 133–136.

[4] L.-S. Huang and C.-H. Yang, “A Novel Approach to Robust Speech Endpoint Detection in Car Environments,” in *ICASSP’00*, Istambul, Turkey, May 2000, vol. 3, pp. 1751–1754.

[5] L.R. Rabiner, C.E. Schmidt, and B.S. Atal, “Evaluation of a Statistical Approach to Voiced-Unvoiced-Silence Analysis for Telephone-Quality Speech,” *THE BELL SYSTEM TECHNICAL JOURNAL*, vol. 56, no. 3, pp. 455–482, Mar. 1977.

[6] W.-H. Shin, B.-S. Lee, Y.-K. Lee, and J.-S. Lee, “Speech/Non-Speech Classification Using Multiple Features for Robust Endpoint Detection,” in *ICASSP’00*, Istambul, Turkey, May 2000, vol. 3, pp. 1399–1402.

[7] C. Mokbel, L. Mauuary, L. Karray, D. Jouvet, J. Monné, J. Simonin, and K. Bartkova, “Towards improving ASR robustness for PSN and GSM telephone applications,” *Speech communication*, vol. 3, pp. 141–159, May 1997.