Robust Speech/Non-Speech Detection using LDA applied to MFCC for Continuous Speech Recognition

Arnaud Martin, Géraldine Damnati and Laurent Mauuary

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

Abstract

Continuous speech recognition applications need precise detection because the number of words to recognize is unknown and vocabulary words can be short. The speech/non-speech detection must be robust to the boundary precision. In this work, a new approach to evaluate detection algorithm for continuous speech recognition is presented. The speech/non-speech detection using energy parameter combined with a Linear Discriminant Analysis (LDA) applied to Mel Frequency Cepstrum Coefficients (MFCC) is compared to the algorithm based on signal to noise ratio (SNR). The LDA applied to MFCC for speech/non-speech detection improves recognition performance in noisy environment and for continuous speech recognition applications.

1. Introduction

The continuous speech recognition performance decreases in part due to imperfect speech detection like for isolated word recognition in a very noisy environment, but not only in this case. Indeed, continuous speech recognition applications need a precise detection because for these applications there is a problematic rejection model, some vocabulary words are short, and the number of words to recognized is unknown, whereas the usual word recognition applications. Therefore efficient speech/non-speech detection is crucial.

Few speech/non-speech detection systems for continuous speech recognition are described in literature. The implications of a word boundary detection using pitch variation are discussed in [1]. The moraic fundamental frequency is used in [2] to improve word boundary detection for continuous speech recognition. Several speech/non-speech detection systems use energy with other parameters to improve speech/non-speech detection in noisy environment. The classification and regression tree method is used like a data fusion method in [3] for word recognition in noisy environment.

In [4], in order to improve recognition performance in noisy environment, we use energy with the LDA applied to MFCC, computed in several recognition systems like in [5]. In the case of two classes, the non-speech and speech classes, a LDA applied to MFCC determines a linear function to integrate all MFCC like a single coefficient.

In this work, the LDA applied on MFCC speech/non-speech detection presented in [4] is used to improve the SNR-based algorithm for continuous speech recognition. This speech/non-speech detection is compared to the algorithm based on SNR [6]. In order to evaluate the speech/non-speech detection we follow the two steps evaluation presented in [7]. In first step, evaluation is made in terms of detection errors, and in second step in terms of recognition errors, established here for continuous speech recognition.

This paper is organized as follows: section 2 recalls the two previous algorithms based on SNR and based on LDA applied to MFCC. Section 3 presents the detection system evaluation principles for continuous speech recognition, and performance of both algorithms. Finally, we conclude in section 4.

2. Algorithm Description

The speech/non-speech detection algorithms are based on an adaptive five state 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 the frame energy and some duration constraints (see Figure 1). Hence one energy condition C1 and two duration conditions C2 and C3 are considered. Duration conditions are controlled with 6 actions.

The two states: *plosive or silence and possible speech continuation* are introduced in order to cope with the energy variability in the observed speech. In the case of continuous speech detection, the between-word silence is longer that the between-phonemes silence. Hence the silence duration (SiD) threshold is changed from 240ms to 960ms. In order not to have a too long silence at the end of detection, the end of detection is 720ms before.

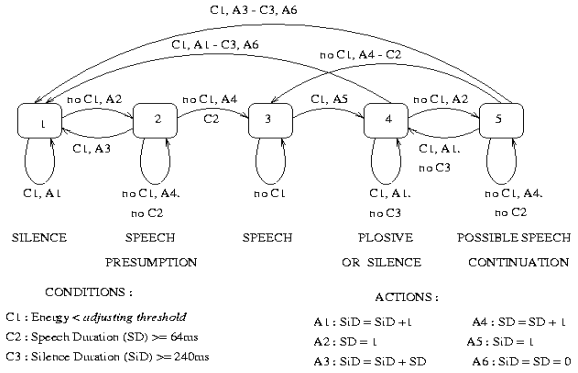

Figure 1. Five state automaton.

2.1. SNR criterion

The SNR criterion uses the estimations for long-term and short-term signal energy. The long-term energy estimate

(LTEE) is recursively computed with the short-term energy (STE) by:

$$LTEE(n+1) = LTEE(n) + (1-\lambda)(STE(n) - LTEE(n)), \quad (1)$$

where λ, is the forgetting factor, close to 1.
The condition C1 is defined by the comparison of the short term and long-term signal energy difference in dB, to an adaptive threshold [6], referred as *adjusting threshold*.

2.2. LDA criterion

The LDA criterion is based on the SNR criterion [4]. The LDA discriminates two classes, the noise/non-speech class and the speech class. The principle is to find a linear function $\boldsymbol{a}$ maximizing between-class variance and minimizing within-class variance. The between-class covariance matrix is noted $\boldsymbol{E}$, and the global covariance matrix $\boldsymbol{T}$. In two classes case, $\boldsymbol{E}$ is such as:

$$\boldsymbol{E} = \boldsymbol{c}\boldsymbol{c}^*, \quad (2)$$

and $\boldsymbol{a}$ is such as:

$$\boldsymbol{a} = \boldsymbol{T}^{-1}\boldsymbol{c}, \quad (3)$$

with $\boldsymbol{a}^*\boldsymbol{T}^{-1}\boldsymbol{a} = 1$, and $\boldsymbol{c} = \frac{\sqrt{n_n n_s}}{n_n + n_s}(\bar{\boldsymbol{x}}_n - \bar{\boldsymbol{x}}_s)$, where n_n is the number of noise frames, n_s the number of speech frames, $\bar{\boldsymbol{x}}_n$ is noise MFCC mean, and $\bar{\boldsymbol{x}}_s$ is speech MFCC mean.

The linear function, obtained by LDA applied to MFCC, is calculated using two learning databases described in section 3.1. To decrease the number of false detections, another condition is added: C4, see Figure 2. When automaton is in *speech presumption* state, if energy is greater than *adjusting threshold*, speech duration is greater than 64ms, and MFCC linear combination, obtained by LDA, is greater than a new threshold, referred as *LDA threshold*, the automaton goes in the *speech* state. If the condition C1 or C4 is realized, the automaton moves to the *silence* state.

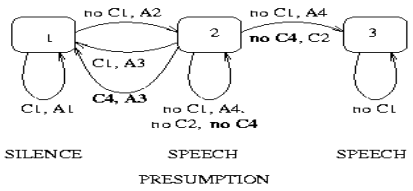

Figure 2. The three first states of the five state automaton with the new condition C4.

This new condition avoids the automaton to go in the *speech* state, when the energy increases due to noise.

3. Evaluations

In order to evaluate the speech/non-speech system, a database that contains French utterances is used. The linear function by LDA applied to MFCC, is calculated using two learning databases that contain isolated French words. Hence the learning is very different with the continuous speech recognition application. The learning databases described in [4], are used to obtain a speech/non-speech detection robust for isolated words recognition in noisy environment and for continuous speech recognition with the same linear function. Both classes, non-speech class and speech class, are determined by the manual segmentation of the two learning databases. The noise segments constitute the noise/non-speech class, and the vocabulary and out-of-vocabulary words segments constitute the speech class.
To evaluate the detection system, LDA criterion performance are compared to those of SNR criterion performance. We follow the evaluation principle presented in [7]. Evaluation is made first in terms of detection errors, and then in terms of recognition errors. The continuous speech recognition errors are not the same as isolated words recognition errors. The recognition errors evaluation according to detection errors is established in this work.
First the databases are described. Next the detection evaluation and the recognition evaluation using the detection system are presented.

3.1. Databases

Two learning databases are used to compute the linear function by LDA applied to MFCC. The learning databases are databases used for French word recognition. The first database includes 1000 phone calls to an interactive voice response service giving movie programs. It was recorded on PSN (Public Switched Network). The corpus contains 25 different vocabulary words. The second learning database is a laboratory GSM (Global System Mobile) database consisting of 51 vocabulary 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
One field database, recorded over PSN, is used to evaluate the speech/non-speech detection system for continuous speech recognition. This database contains 98 phone calls to an interactive spoken dialogue service. Manual segmentation gives 71% of speech segments and 29% of non-speech segments. The speech segments contain 12635 French word occurrences in 2520 utterances, with 1633 vocabulary words.

3.2. Detection Evaluation

First to evaluate the detection system, automatic speech segment detection is compared to manual segmentation of speech and noise periods [7]. To count errors, a two steps procedure is used. Firstly, manual and automatic segments are tied if their temporal intersection exceeds half the duration of the shortest segment. Secondly, errors resulting from the comparison of tied segments are counted. Hence different errors are considered:

- Omission: an utterance is not detected,
- Insertion: a noise (or silence) segment is detected, as speech,
- Regrouping: several utterances are detected as one,
- Fragmentation: one utterance is detected as several.

The insertions can be rejected by the recognition system. These errors are called *recoverable errors*. The other errors unavoidably produce recognition errors. These errors are called *definitive errors*. The recoverable and definitive error rates are calculated with respect to the total number of speech segments.
To compare SNR criterion and LDA criterion, definitive errors according to recoverable errors are plotted varying the *adjusting threshold*. The *LDA threshold* is fixed and was optimized using the two learning databases.

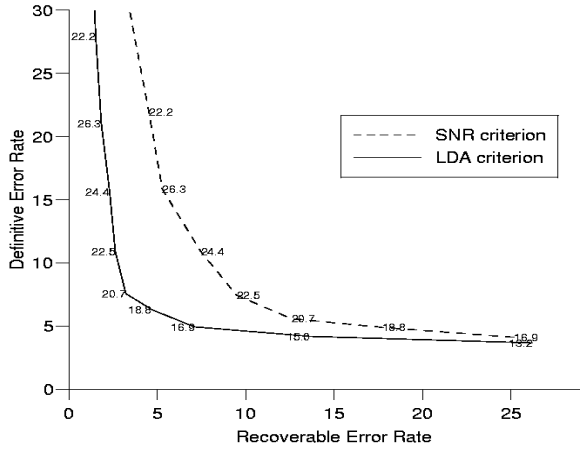

Figure 3. Detection test.

Figure 3 presents detection results for both criteria, for different *adjusting thresholds*. It shows that for one given threshold (for example 16.9dB) LDA criterion produces less recoverable error with the same definitive error rate, which was expected with the new condition C4. The improvement is statistically significant.

But this representation does not show the boundary precision of the detection. Indeed, the speech/non-speech detection does not need precise boundary, for isolated word recognition application, that uses long vocabulary words. But continuous speech recognition needs precise detection because some words are short, and the number of words to recognize in the utterance is unknown.

In Figure 4, the histogram of the number of segments according to the left and right boundary position compared to the left and right manual boundary (frame 0) is plotted, for one fixed *adjusting threshold*, 16.9dB. For example, the marked point A shows that there are about 75% of segments, with the left boundary detected after the manual left boundary (frame 0) or before 5 frames.

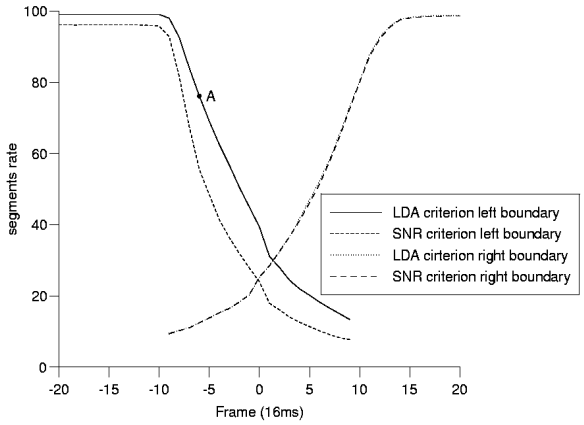

Figure 4. Boundary position according to the manual left and right boundary (frame 0).

Figure 4 shows that the LDA criterion detects right boundary like the SNR criterion, but detects left boundary later that SNR criterion. Hence there is less left widened detections but more left truncated segments. Left boundaries difference come from the integration of the new condition C4.

3.3. Recognition Evaluation

Recognition experiments were conducted using an HMM-based speech recognition system [5]. The used model is a context dependent multigaussian model, and contains 1633 vocabulary words. Insertion of segments can be rejected with a noise-rejected model. Recognition evaluation is made with the speech/non-speech detection. The difference with the usual continuous speech recognition evaluation is that the reference: the manually segmented utterance boundaries can be different from the test segment boundaries. Hence a temporal difference between reference and test segments is possible. Four error types are considered:

- False rejection: one utterance is rejected by recognition system, or not detected. This error is counted in words omission,
- Omission: one word is omitted in the utterance,
- Insertion: one word is added in the utterance,
- Substitution: one word is recognized as another vocabulary word.

First, Figure 5 presents global recognition errors for SNR and LDA criteria and for manual segmentation. Error rates are calculated with respect to total words number. Omission, insertion and substitution rates are represented according to the false rejection rate. Curves are obtained by varying the rejection threshold, and for the *adjusting threshold* giving the minimum recognition errors for each criterion (18.8dB for SNR criterion and 15.0dB for LDA criterion).

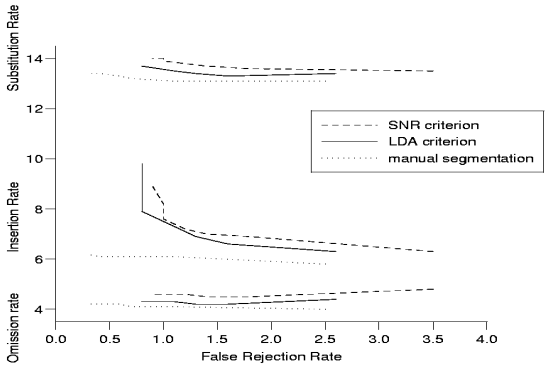

Figure 5. Recognition test.

Figure 5 shows that LDA criterion outperforms the SNR criterion. The improvement on each error is not statistically significant, but on the global error, the improvement is statistically significant. The LDA criterion improvement compared to the SNR criterion is statistically significant (3.75% relative to the global error rate for a fixed rejection threshold). Results on manual segmentation show that further improvements can be achieved, up to 5% relative improvement on the global error rate for a fixed rejection threshold.

In order to identify potential improvements, recognition errors according to detection results are represented in Figure 6 for a fixed rejection threshold and optimized *adjusting threshold*. For these thresholds, the global error rate is 26.7% for SNR criterion, 25.7% for LDA criterion and 24.4% for the manual segmentation. Correct, regrouping, fragmentation, omission and insertion segments are distinguished for detection (number of each is indicated on the figure). Error rates are still calculated with respect to the total number of words, a logarithm scale is used.

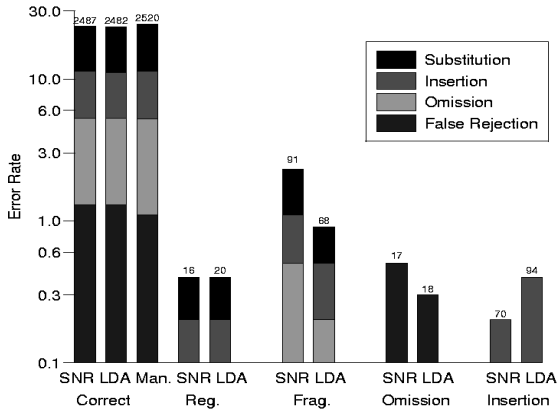

Figure 6. Recognition errors according to detection results.

Figure 6 shows that most of the recognition errors happen on correct detections. Manual segmentation gives more global error rate that errors happening on correct detection for LDA and SNR criteria, but gives less error rate proportionally to the correct detection (all detections are correct for the manual segmentation). Proportionally, error rates are more important on detection errors than for correct detections. The LDA criterion improvement compared to the SNR criterion comes essentially from the recognition errors on the fragmentation detections (43% relative to the global error rate coming from fragmentation detections). Indeed LDA criterion gives less fragmentation segments that SNR criterion.

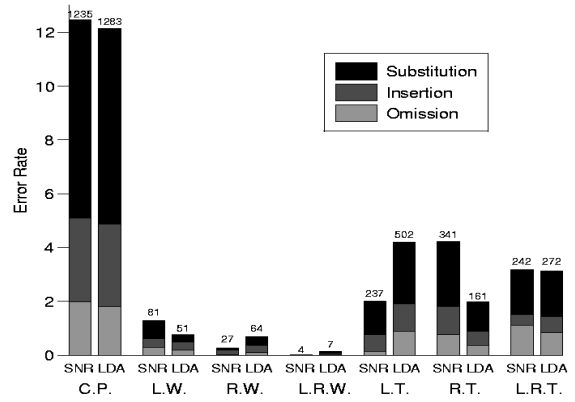

Figure 7. Recognition error on correct detection according to the boundaries position.

On Figure 4 differences between LDA and SNR criteria on left boundaries positions appear. Figure 7 presents the recognition error rates on correct detection according to the boundary positions. The position is considered correct if the manual segment on the left and right boundaries is not truncated and not wider than 160ms on the left boundary, and than 240ms on right boundary (according to the automaton parameters). The segments are widened (W.) if the detection is not correct and the manual segments are not truncated. Correct position is referred C.P., a left truncated segment, L.T., a right truncated segment R.T., and L.R.T. a left and right truncated segment. A left widened segment is referred L.W., a right widened segment, R.W., and a left and right widened segment L.R.W. Number of each case is indicated. Error rates are calculated with respect to total number of words, for the optimized *adjusting threshold* and a fixed rejection threshold, explaining the difference on the right boundary.

Recognition errors for both LDA and SNR criteria are more important on truncated segments than on widened boundaries. LDA criterion gives more errors for left truncated segments and right widened boundaries, but less errors for right truncated segments and left widened boundaries. However the difference is not significant. Error rates are proportionally more important on truncated segments than on correct detections. Hence truncated segments must be reduced to decrease global error rates.

4. Conclusions

This work presents an evaluation of speech/non-speech detection for continuous speech recognition. The evaluation of the speech/non-speech detection used the LDA applied to MFCC gives significant improvement for detection results and for global recognition performance, 3.75% relative improvement.

The comparison with recognition results on manual segmentation and LDA criterion shows that a 5% relative improvement is potentially possible, on the global error rates. This study shows that the truncated segments and detection errors (regrouping, fragmentation, omission, insertion) lead 1.3% global recognition errors. To improve the speech/non-speech detection, these errors must be reduced.

The integration of the LDA applied to MFCC in the detection system outperforms the SNR criterion in noisy environment [4] and for continuous speech recognition.

To reduce detection errors, we are investigating another combination of the linear function calculated by LDA applied to MFCC and the energy.

5. References

[1] Ramana Rao, G.V. and Srichand, J., "Word Boundary Detection using Pitch Variations", ICSLP, Philadelphia, USA, Vol. 2, pp. 813-816, 1996.

[2] Iwano, I. And Hirose, K., "Prosodic Word Boundary Detection Using Statistical Modeling of Moraic Fundamental Frequency Contours and its Use for Continuous Speech Recognition", ICASSP, Phoenix, USA, Vol. 1., pp. 133-136, 1999.

[3] Shin, W.-H., Lee, B.-S., Lee, Y.-K., Lee, J.-S., "Speech/Non-Speech Classification Using Multiple Features for Robust Endpoint Detection", ICASSP, Istanbul, Turkey, Vol. 3, pp. 1399-1402, 2000.

[4] Martin, A., Charlet, D. and Mauuary, L., "Robust Speech/Non-Speech Detection using LDA applied to MFCC", *ICASSP*, Salt Lake City, USA, 2001.

[5] Mokbel, C., Mauuary, L., Karray, L., Jouvet, D., Monné', J., Simonin, J., Bartkova, K., "Towards improving ASR robustness for PSN and GSM telephone applications", Speech communication, Vol. 3, pp.141-159, 1997.

[6] Mauuary, L. and Monné, J., "Speech/Non-Speech Detection for Voice Response Systems", Eurospeech, Berlin, Germany, pp. 1097-1100, 1993.

[7] Mauuary, L. and Karray, L., "The Tuning of Speech Detection in the Context of a Global Evaluation of Voice Response System", Eurospeech, Rhodes, Greece, Vol. 3, pp. 1539-1542, 1997.