Voicing Parameter and Energy Based Speech/Non-Speech Detection for Speech Recognition in Adverse Conditions

Arnaud Martin[1], Laurent Mauuary[2]

[1]Université de Bretagne Sud, Valoria
56000 Vannes - France
arnaud.martin@univ-ubs.fr
[2]France Télécom R&D, DIH/IPS
2, av. P. Marzin, 22307 Lannion Cedex - France

Abstract

In adverse conditions, the speech recognition performances decrease in part due to imperfect speech/non-speech detection. In this paper, a new combination of voicing parameter and energy for speech/non-speech detection is described. This combination avoids especially the noise detections in real life very noisy environments and provides better performances for continuous speech recognition. This new speech/non-speech detection approach outperforms both noise statistical based [1] and Linear Discriminate Analysis (LDA) based [2] criteria in noisy environments and for continuous speech recognition applications.

1. Introduction

In adverse conditions, the speech recognition performances decrease in part due to imperfect speech/non-speech detection. Efficient speech/non-speech detection is crucial, on the hand in noisy environments and on the other hand for continuous speech recognition. Indeed, in very noisy environments, the speech/non-speech detection may indicate noises as speech to the speech recognition system, producing many errors. It is also critical for continuous speech recognition systems. The out of vocabulary words rejection is a very difficult task because: some vocabulary words are short. Moreover, the number of words to recognize in a sentence is unknown, unlike the usual isolated word recognition applications.

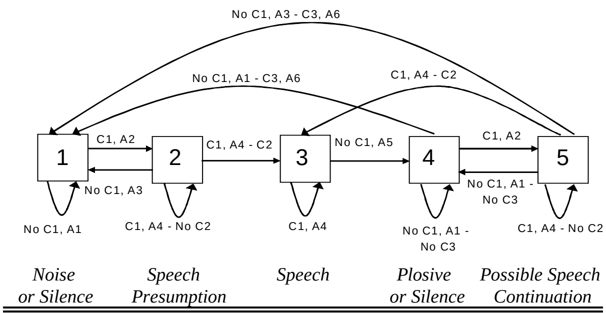
The most widely used parameter for speech/non-speech detection systems is energy. This single parameter is not sufficient in noisy environment. In order to discriminate the noise and speech signal, several studies use the energy with a voicing parameter. Indeed, voiced sounds are a characteristic of speech. In the acoustic domain, a voicing parameter can be determinate by studying the variations of the fundamental frequency, referred to as F_0.

In order to estimate a voicing parameter, a zero crossing rate can be calculated and used with the energy ([3], and [4]). However, the zero crossing rates are too unstable in noisy environments [5]. Hence, a precise F_0 estimation must be calculated, in order to calculate a precise voicing parameter. Many studies propose an energy-voicing parameter combination (with or without other parameters) for all the frames like in [6], and [7]. However the energy is a good parameter when the Signal-to-Noise-Ratio (SNR) is high enough. Therefore, we propose a new energy-voicing parameter combination, only for energetic frames, in order to discriminate energetic noise and speech frames.

This paper is organized as follows: section 2 recalls our previous work: both noise statistical based and LDA based criteria for speech/non-speech detection. Section 3 presents the used F_0 estimation and how the new energy-voicing parameter combination is achieved. Finally, section 4 describes the evaluation of this new criterion.

2. Previous Criteria

All speech/non-speech detection can be seen as an automaton, with 2 states (speech/non-speech) or more states. Our previous studies show that the adaptive five state automaton gives very good performances [2]. The five states are: *noise or 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 Fig. 1.



Conditions	Actions
C1: Energy>*adjusting threshold*	A1: SiD=SiD+1
C2: Speech Duration (SD)≥64ms	A2: SD=1
C3: Silence Duration (SiD)≥240ms	A3: SiD=SiD+SD
	A4: SD=SD+1
	A5: SiD=1
	A6: SiD=SD=0

Figure 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 (with-word silence) and to avoid various kind of noise. Hence, the *speech presumption* state prevents 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 transition from one state to another can be controlled by different C1-condition. We present here both best criteria until now.

2.1. Noise Statistical Criterion

The noise energy distribution is assumed a normal distribution (μ, σ^2) [1]. The noise energy mean and standard deviation are estimated recursively in the *noise or silence* state by:

$$\hat{\mu}(n) = \hat{\mu}(n-1) + (1-\lambda)(E(n) - \hat{\mu}(n-1)), \quad (1)$$

and

$$\hat{\sigma}(n) = \hat{\sigma}(n-1) + (1-\lambda)(|E(n) - \hat{\mu}(n-1)| - \hat{\sigma}(n-1)), \quad (2)$$

where n is the current frame, $E(n)$ the energy, and λ is a forgetting factor optimized to 0.99 in (1) and to 0.95 in (2). For a given frame, noise (or non-speech) frame is tested, comparing the centered and normalized energy of the frame $r_{NS}(E(n)) = (E(n) - \hat{\mu}(n))/\hat{\sigma}(n)$ to an *adjusting threshold*. Hence the condition C1 is given by:

$$\text{C1: } r_{NS}(E(n)) > \textit{adjusting threshold}. \quad (3)$$

This criterion is referred to as the NS criterion [1].

2.2. LDA Criterion

This method discriminates two classes, the noise class and the speech class. The idea is to find a linear function a that maximizes between-class variance and minimizes within-class variance.

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

$$a^*\mathrm{T}a = a^*\mathrm{D}a + a^*\mathrm{E}a. \quad (4)$$

So the linear function a is such as $a^*\mathrm{D}a$ is minimal and $a^*\mathrm{E}a$ is maximal. We have to solve:

$$\mathrm{T}^{-1}\mathrm{E}a = \lambda a, \quad (5)$$

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

$$\mathrm{E} = cc^*, \quad (6)$$

with

$$c_j = \frac{\sqrt{n_n n_s}}{n_n + n_s}(\bar{x}_{nj} - \bar{x}_{sj}), \quad (7)$$

where n_n is the number of noise frames, n_s the number of speech frames, $\bar{x}_{nj}$ is noise j^{th} MFCC mean, and $\bar{x}_{sj}$ is speech j^{th} MFCC mean. Hence the equation (5) gives $a=\mathrm{T}^{-1}c$, the only linear function.

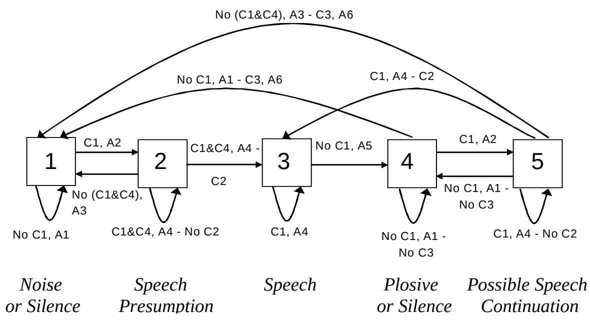


Figure 2: Five State Automaton with a new condition C4.

The linear function a is calculated on two learning databases (described in section 4.1) using the Mel Frequency Cepstrum Coefficients (MFCCs). This linear function is integrated with the condition C1 of the NS criterion using an additional condition in the automaton (see Fig. 2), referred to as C4, given by:

$$\text{C4: } a.\mathrm{X}(n) < \text{LDA threshold}, \quad (8)$$

where X(n) is the MFCCs vector of the frame n, and LDA threshold is optimized on both learning databases.

This condition C4 is added between the speech presumption and speech state in order to decrease the false detections of noises. This criterion is referred to as the NS+LDA criterion [2].

3. New Energy-Voicing Combination

In order to obtain a voicing parameter, a precise F_0 estimation is calculated. The F_0 estimation introduced in [8] is computed on the entire signal (voiced and unvoiced sound). The signal harmonicity is calculated by intercorrelation with a comb-function.

Hence, a F_0 value is obtained every 4 ms (4 values for each 16 ms frame). In order to avoid the artifacts the median is calculated, referred to as *med*:

$$med(n) = med(F_0(n-1), F_0(n), F_0(n+1)), \quad (9)$$

where n is the current sub-frame of 4 ms. Then, a mean-variation, referred to as $\overline{\delta med}$, is calculated over N sub-frames:

$$\overline{\delta med}(n) = [1/N] \sum_{m=n-N}^{n} |med(m) - med(m-1)| \cdot \quad (10)$$

This mean-variation is used as an estimation of a voicing parameter. A new condition C4 is defined by this voicing parameter compared to a threshold. It is integrated with the condition C1 of the NS criterion between the *speech presumption* and *speech* state in order to decrease the false detection of noises; like in the NS+LDA criterion (see Fig. 2). C4 is given by:

$$\text{C4: } \overline{\delta med}(4m) < \textit{VP threshold}, m \in \mathbb{N}^*. \quad (11)$$

The *VP threshold* is optimized on both learning databases. In order to obtain a decision each 16 ms frame, the mean-variation is considered every $4m$ sub-frames. When the new automaton (described on Fig. 2) is in *speech presumption* state, if the energy is high enough (C1 is realized), speech duration is greater to 64 ms (C2 is realized), and the frame is voiced (C4 is realized), the automaton goes in the *speech* state. Hence, the condition C4 prevents the automaton from going in the *speech* state for energetic noises, so the noise detections will decrease. This new criterion is referred to as the NS+VP criterion.

4. Experiments

Evaluations has been carried out on two databases. The first one contains a lot of real life noises, and the other one is a continuous speech database. In the case of continuous speech, the between-word silence is longer that the between-phonemes silence. Hence the silence duration (SiD) threshold in the condition C3 of the automaton is changed from 240 ms to 960 ms. In order not to have a too long silence at the end of the detection, the end of detection is 720 ms before (described in [9]). Evaluations are made, first in terms of detection errors, and then in terms of recognition errors.

4.1. Databases

Two learning databases are used to optimize thresholds and to compute the linear function by LDA. The first database includes 1000 phone calls to an interactive voice response service, which was recorded on PSN (Public Switched

Network). This corpus contains 25 different French vocabulary words. The second learning database is a laboratory GSM (Global System Mobile) database consisting of 51 French vocabulary words, including 390 phone calls.

Another laboratory GSM database, referred to as GSM database, is used for evaluations. It contains 65 French vocabulary words, including 390 phone calls from different environments: indoor, outdoor, stopped car, and running car. In order to study criteria according to the noise level, the database is divided into two parts: the first part with SNR inferior to 18 dB, and second part with SNR superior to 18 dB. Manual segmentation gives 85% of vocabulary word segments, 3% of out-of-vocabulary word segments, and 11% of noise segments.

One field database, recorded over PSN, is used to evaluate criteria for continuous speech recognition applications. This database, referred to as continuous PSN 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

4.2. Detection experiments

To evaluate speech/non-speech detection in terms of detection errors, automatic speech segment detection is compared to manual segmentation of speech and noise periods. Hence, different error types are considered: omission (a vocabulary or out-of-vocabulary word is not detected), insertion (a noise is detected as speech), regrouping (several words are detected as one), and fragmentation (on word is detected as several).

Noise detections can be rejected by the rejection model of the recognition system. These errors are called *recoverable errors*. The omission, regrouping, and fragmentation errors, unavoidably producing recognition errors, are called *definitive errors*. Recoverable and definitive error rates are calculated with respect to the total number of speech segments. To compare the three criteria, definitive errors according to recoverable errors are plotted for different *adjusting thresholds.*

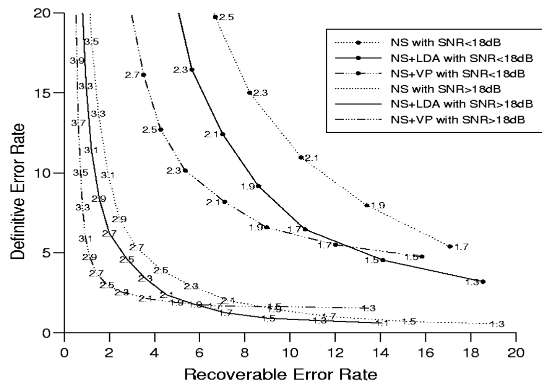


Figure 3: Detection test on GSM database according to the SNR.

Fig. 3 shows the detection performances for the NS, NS+LDA and NS+VP criteria on the GSM database according to the SNR. The *adjusting thresholds* are noted on the curves. The NS+VP criterion outperforms both NS and NS+LDA criteria. The improvement is statistically significant on both database parts. For one observed threshold (e.g. 1.9 on the database part with SNR inferior to 18 dB) we note the same recoverable error reduction than the NS+LDA criterion (expected by the condition C4), but we observe too a definitive error reduction.

Detection results for the continuous speech recognition application are presented on Fig. 4. Here also, the NS+VP criterion outperforms both NS and NS+LDA criteria. However both NS+LDA and NS+VP criteria results are very close. The improvement of both criteria is due to the recoverable error rate reduction, and is statistically significant.

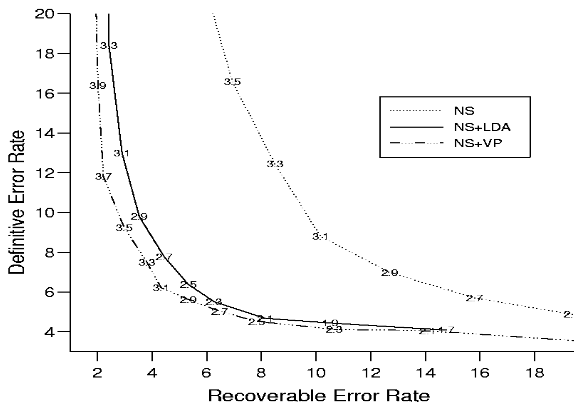


Figure 4: Detection test on continuous PSN database.

4.3. Recognition experiments

Recognition experiments were conducted using an Hidden Markov Model-based speech recognition system [10]. The used model is a context dependent multigaussian model, and contains 65 vocabulary words for the isolated word recognition and 1633 for the continuous speech recognition. Insertion of segments can be rejected with a noise-rejected model. Recognition evaluation is made with the speech/non-speech detection results obtained with the *adjusting threshold* giving the minimum recognition errors. Curves are obtained by varying the rejection threshold. For the isolated word recognition, three errors types are considered: substitution (a vocabulary word is recognized as another vocabulary word), false acceptance (a noise or out-of-vocabulary word is recognized as a vocabulary word), and false rejection (a vocabulary word is rejected, or not detected).

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.

For the continuous speech recognition, the difference with the usual 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 is possible. In this case four error types are considered: substitution (a word is recognized as another vocabulary word), insertion (one word is added in the utterance), omission (one word is omitted in the utterance), and false rejection (one utterance is rejected by recognition system, or not detected). This error is counted in terms of words omitted. Error rates are calculated with respect to the total words number in the database.

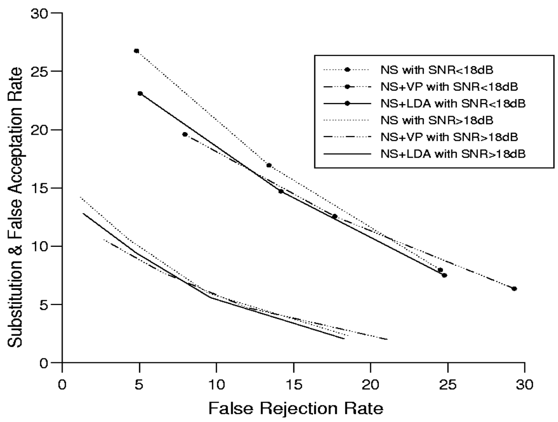


Figure 5: Recognition test on GSM database according to the SNR.

Fig. 5 presents recognition results of the three criteria on the GSM database according to the SNR. Notice that both NS+LDA and NS+VP criteria performances are very close on both database parts. The improvement compared to NS criterion is statistically significant for a false rejection rate inferior to 10 % (generally considered as a maximum value for the user). However the improvement is not statistically significant for SNR superior to 18 dB. NS+VP criterion reduces noise detections, and therefore allows great improvements for the speech/non-speech detection performances. But noise detections can be rejected by the rejection model and that do not reduce error rates of the speech recognition system. However the detection computational cost is smaller than the rejection computational cost.

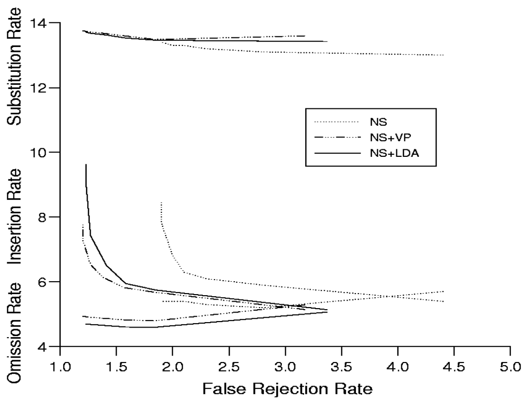


Figure 6: Recognition on continuous PSN database.

Fig. 6 shows the continuous speech recognition performances for the three criteria of the speech/non-speech detection. Both NS+VP and NS+LDA criteria results are very close and are better than NS criterion results. The improvement on the global errors is statistically significant. The improvement is due to the insertion and omission error rates reduction.

5. Conclusions

This work presents a new speech/non-speech detection based on energy-voicing parameter combination. This combination made for energetic frames provides significant improvements in adverse conditions (noisy environments and for continuous speech applications). The NS+VP criterion results in less noise detections that do not allow a reduction of recognition error rates if the rejection model is efficient. However the NS+VP computational cost is inferior to rejection computational cost at the recognition system level. This new criterion outperforms both NS and NS+LDA criteria and provides significant improvements.

6. References

[1] L., Karray, and J., Monné, "Robust Speech/Non-Speech Detection in Adverse Conditions Based on Speech Statistics," in *ICSLP*, Dec. 1998, Australia, vol. 4, pp. 1471-1474.

[2] A., Martin, D., Charlet, and M., Mauuary, "Robust Speech/Non-Speech Detection Using LDA Applied to MFCC," in *ICASSP*, May 2001, USA, vol. 1, pp. 237-240.

[3] M.H., Savoji, "A Robust Algorithm for Accurate Endpointing of Speech Signals," *Speech Communication*, vol. 8, pp.45-60, 1989.

[4] A., Ganapathiraju, L., Webster, J., Trimble, K., Bush, P., Korman, "Comparison of Energy-Based Endpoint Detection for Speech Signal Processing," in *IEEE Southeastcon*, USA, Apr. 1996, pp. 500-503.

[5] L.-S., Huang, C.-H., Yang, "A Novel Approach to Robust Speech Endpoint Detection in Car Environments," in *ICASSP*, May 2000, Turkey, vol. 3, pp. 1751-1754.

[6] H., Kobatake, K., Tawa, A., Ishida, "Speech/Non-Speech Discrimination for Speech Recognition System under Real Life Noise Environments," in *ICASSP*, United-Kingdom, May 1989, vol. 1, pp. 365-368.

[7] G.V., Ramana Rao, J., Srichand, "Word Boundary Detection Using Pitch Variations," in *ICASSP*, USA, Oct. 1996, vol. 2, pp. 813-816.

[8] P., Martin, "Comparison of speech detection by cepstrum and combination analysis," in *ICASSP*, 1982, pp. 180-183.

[9] A., Martin, G., Damnati, L., Mauuary, "Robust Speech/Non-Speech Detection Using LDA Applied to MFCC for Continuous Speech Recognition," in *Eurospeech*, Sep. 2001, Denmark, vol. 2, pp. 885-888.

[10] C., Mokbel, L., Mauuary, L.,Karray, D., Jouvet, J., Monné, J., Simonin, K., Barrtkova, "Towards improving ASR robustness for PSN and GSM telephone applications," *Speech Communication*, vol. 3, pp. 141-159, May 1997.