IDIAP RESEARCH REPORT

Exploiting temporal context for speech/non-speech detection

Sree Hari Krishnan Parthasarathi,
Petr Motlicek, and
Hynek Hermansky *

IDIAP–RR 08-21

April 2008

* IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

Rapport de recherche de l'IDIAP 08-21

Exploiting temporal context for speech/non-speech detection

Sree Hari Krishnan Parthasarathi,
Petr Motlicek, and
Hynek Hermansky

April 2008

Résumé. In this paper, we investigate the effect of temporal context for speech/non-speech detection (SND). It is shown that even a simple feature such as full-band energy, when employed with a large-enough context, shows promise for further investigation. Experimental evaluations on the test data set, with a state-of-the-art multi-layer perceptron based SND system and a simple energy threshold based SND method, using the F-measure, show an absolute performance gain of 4.4% and 5.4% respectively, when used with a context of 1000 ms. ROC based performance evaluation also reveals promising performance for the proposed method, particularly in low SNR conditions.

1 Introduction

The primary objective of our work is to design a simple speech/non-speech detection (SND) algorithm that can be implemented on low power devices. While accurate determination of SND boundaries in difficult environments is still a challenging task using sophisticated algorithms, performing it using simple algorithms is still an unsolved problem.

Design of features is an important problem for SND. Traditionally SND algorithms use short-term features. Short-term features are susceptible to changes in background noise or variations in the speech signal. The importance of longer temporal context is well known for Automatic Speech Recognition (ASR) [5, 4]. Two recent studies of SND, [6] and [7], exploit long temporal context using modulation spectrum.

In this paper, we design a simple algorithm that exploits the temporal context of logarithmic full-band energy. For this, the weights of the context around the frame-to-be-classified, need to be obtained. This is obtained using the Linear Discriminant Analysis (LDA). Further, this method gives us an interpretation in terms of a filter in the modulation spectral domain.

The rest of the paper is organized as follows. In Section 2, a brief overview of LDA is provided. Next, Section 3 discusses the proposed method in detail. Description of the experimental evaluation and the data set is provided in Section 4. Finally, we draw some conclusions in Section 5.

2 Review of linear discriminant analysis

LDA [2] is a linear transformation that reduces the dimensionality of the data in such a way that the information important for classification is preserved. For this reduced subspace, it yields a set of linearly independent bases. In a k-class classification problem, the number of bases is equal to $k-1$. In the discussion that follows, an overview of LDA is provided for a two-class problem.

Let $\{\mathbf{x}_i{}^k\}$ denote a set of d-dimensional feature vectors, where $\mathbf{x}_i{}^k$ represents the i^{th} example of the k^{th} class, where $k=1,2$. The number of examples for each class is denoted by n_k. Let $\mathbf{m}_k = \sum_{i=1}^{n_k} \mathbf{x}_i{}^k$ denote the mean vectors of the respective classes. Further, let us denote $\mathbf{m}$ as the mean of the entire data set. Following [2], we define the within-class (s_w) and between-class (s_b) scatter matrices as follows :

$$
\begin{aligned}
s_k &= \sum_{i=1}^{n_k} (\mathbf{x}_i{}^k - \mathbf{m}_k)(\mathbf{x}_i{}^k - \mathbf{m}_k)^t \qquad k \in \{1,2\} \\
s_w &= s_1 + s_2 \\
s_b &= \sum_{k=1}^{2} (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^t. \qquad (1)
\end{aligned}
$$

LDA seeks to project the data onto a weight vector $\mathbf{w}$ such that, in the projected space, the distance between the means of the two classes is maximized while minimizing the within-class scatter s_w. This is formulated as the maximization of the objective function $J(.)$:

$$
J(\mathbf{w}) = \frac{\mathbf{w}^t s_b \mathbf{w}}{\mathbf{w}^t s_w \mathbf{w}}. \qquad (2)
$$

The solution $\mathbf{w}$, is the eigen vector of ${s_w}^{-1} s_b$. For a two-class problem, this simplifies to :

$$
\mathbf{w} = {s_w}^{-1}(\mathbf{m}_1 - \mathbf{m}_2). \qquad (3)
$$

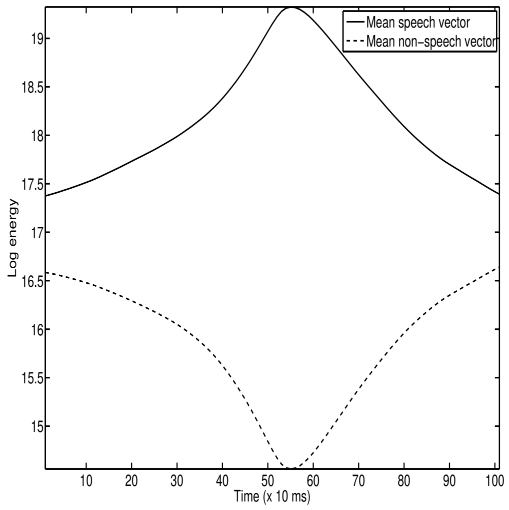

Fig. 1 – *Mean speech and non-speech vectors.*

3 Weighting temporal context for SND

3.1 Features

The first step is to obtain feature vectors, $\mathbf{x}_i{}^k$, for the two classes. This is done as follows : for each speech signal in the data set, the logarithmic full-band energy is computed using a rectangular analysis window of length and shift 25 ms and 10 ms, respectively. Feature vectors are extracted by considering overlapping windows (i.e., shift of 10 ms) on this temporal trajectory. This method of extracting features introduces a context around the frame under consideration.

To better understand the choice of this feature vector, we briefly discuss the characteristics of the speech and non-speech data : the mean speech and non-speech vectors, $\mathbf{m}_1$ and $\mathbf{m}_2$, are shown in Fig. 1. These vectors are 1010 ms long (101 frames at 10 ms frame rate). It can be seen that these vectors are quite distinct for speech and non-speech. Further, these vectors are easily interpretable. Since speech frames have higher energy than non-speech on an average, the mean speech vector shows a pronounced peak at the center. The converse is true for the mean non-speech vector.

3.2 LDA to obtain the weights of the context

In this section, we obtain the weights of the context around the the frame that is to be classified. For this, we use LDA. The training data used for LDA is described in Section 4.1.

3.2.1 Training the weight vector

The label of the class at the center of the feature vector determines the training targets. LDA procedure outlined in section 2 is used to obtain the weight vector for classification.

Fig. 2(a) shows the weight vector (flipped left to right, for interpretation as an impulse response) obtained by LDA using a feature vector of dimension 101. This can be considered as the impulse response of a matched filter. The valley at the center can be understood from the fact that the mean speech and non-speech vectors suggest that the dimensions most important for classification are the 20 frames around the center. This can be interpreted that a context of 200 ms around the center is important for classification. Also, note that from equation 3, reducing the feature vector dimension to

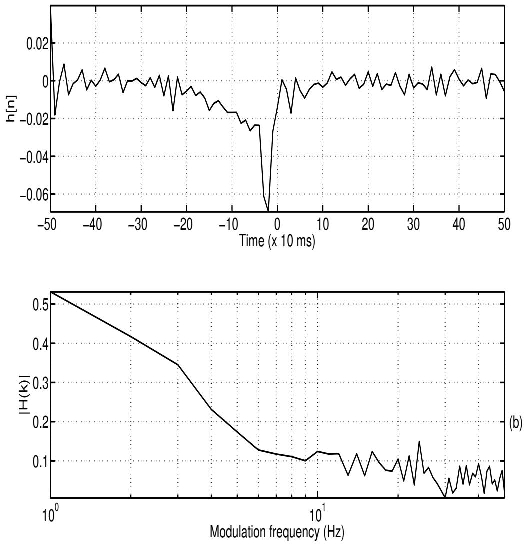

FIG. 2 – *Impulse and magnitude response of the matched filter obtained by LDA.*

one, reduces the method to energy thresholding. Further, it is interesting to observe that the frequency response of the filter shows a cut-off frequency in the modulation frequency domain at 4 to 6 Hz.

3.2.2 Computation of thresholds

The final step in the training procedure is the determination of the optimal threshold (θ). The threshold is computed as follows :

1. Project the training examples on the weight vector.
2. Plot the distribution of projected values for speech and non-speech.
3. Determine the point where the two distributions cross each other, to obtain θ.

3.2.3 Determination of SND boundaries

During testing, the feature vectors ($\mathbf{x}_i$) of the speech signal are computed as described in Section 3.1. The vectors are projected on to $\mathbf{w}$. These projected values are then compared with the threshold (θ), to determine the class :

$$\begin{aligned} \mathbf{w}^t\mathbf{x} &\geq \theta, \text{ assign non-speech} \\ &< \theta, \text{ assign speech.} \end{aligned} \tag{4}$$

4 Experiments and Evaluation

4.1 Experimental data set

Experiments were conducted on a subset of the NIST meeting room corpus [3]. Data obtained from close-talking microphones were used for the experiments. The sampling rate and the quantization of the data are 16 kHz and 16 bits respectively. The training and testing sets consist approximately of 1 and 3 hours of data respectively. The overall ratio of non-speech to speech segments is 46% : 54%. The labels for the training and testing data were obtained by forced-alignment of ASR phoneme models [1]. All phonemes except 'sil' were considered 'speech', while the 'sil' regions were labeled as 'non-speech'.

Since the data used in this study is from close-talking microphones, the signals are relatively clean. To study the effect of noise on the SND systems, babble noise from NOISEX-92 database [8] was added at various SNR levels.

4.2 Methods used for comparison

Since the primary task of the study is to investigate the utility of long term information for SND, we evaluate the proposed method against two short term methods : (a) a state-of-the-art multi-layer perceptron (MLP) based method [1] and (b) a simple short-term energy threshold based method.

The MLP based method uses 12 MF-PLP coefficients along with their first and second derivatives. To these, the following auxiliary features are added : normalized energy from all channels, signal kurtosis, mean cross-correlation and maximum normalized cross-correlation. The MLP is trained on 98 hours of training data, with a hyperbolic tangent hidden activation function and soft-max output activation function.

The energy based method computes short-term log energy and uses a threshold to make the SND decisions.

4.3 Evaluation using ROC curves

In the first set of experiments, a context of 1000 ms is used for the proposed method. We compare the proposed method with the MLP based method at various SNR levels. In the second set of experiments, the length of the context is varied to identify the optimal length for SND. As observed earlier, energy based method is obtained by setting the length of the context to zero. Hence, studying the effect of different contextual lengths inherently includes a comparison with the short-term energy based method as well.

We now present the comparison of the proposed method with the MLP based system using the Receiver Operating Characteristics (ROC) curve method, in clean and noisy environments. To plot the ROC curve, we compute "true speech positives" and "false speech positives" by varying the thresholds of the methods. The noisy data is obtained by adding babble noise from NOISEX-92 database at four different SNR levels : 10 dB, 5dB, 0 dB and -5 dB. The results are shown in Figs. 3 and 4.

Fig. 3 shows that the MLP based method performs better than the proposed method when the environment is relatively less noisy. This is indeed not surprising because the MLP based method is trained on many hours of meeting room data and consequently performs well when the testing conditions match the trained conditions.

On the other hand, Fig. 4 illustrates that the proposed system performs better than the MLP based system in significantly noisy conditions. We attribute the performance of the proposed method to long-term contextual information.

4.4 Evaluation using F-measure

Here we note that the ROC method of evaluating algorithms does not measure the sensitivity of the methods to thresholds. As as illustration, for any SND method, the threshold is set for a particular operating point on the development data. When the testing environment is different from the development environment, the threshold changes. Since the threshold cannot be modified for the test data, we want the performance to remain the same.

To evaluate aspect of the SND algorithms, we utilize F-measure. The F-measure, defined as the harmonic mean of "precision" and "recall" is :

$$\mathrm{F} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

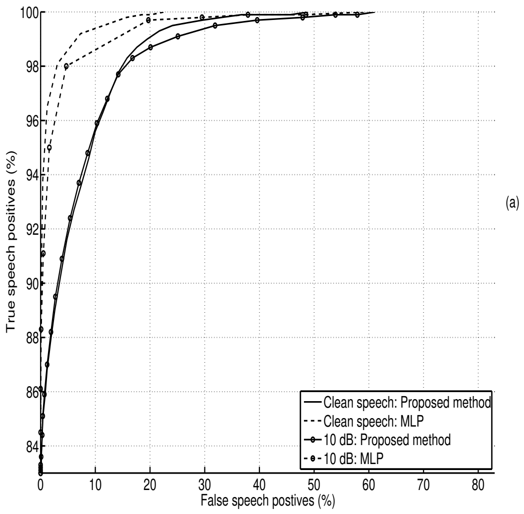

FIG. 3 – *Comparison of proposed and MLP based system in Clean and 10 dB SNR.*

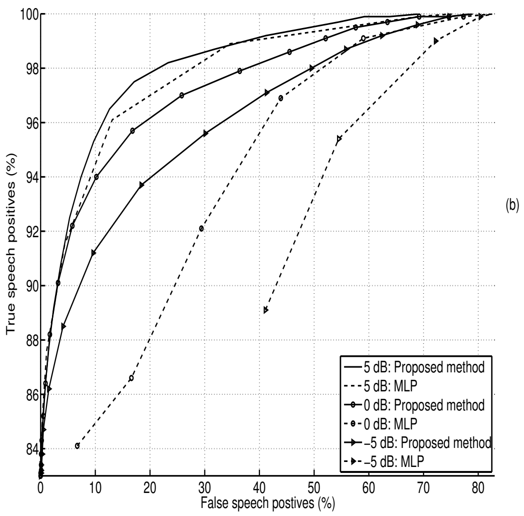

FIG. 4 – *Comparison of proposed and MLP based system in 5, 0 and -5 dB SNR.*

$$\begin{aligned}\text{precision} &= \frac{\text{True positives}}{\text{True positives} + \text{false positives}}\\ \text{recall} &= \frac{\text{True positives}}{\text{True positives} + \text{false negatives}}.\end{aligned} \tag{5}$$

A high value of recall with a high precision, yields a high F-measure. The maximum value of F-measure that can be obtained is 1. This value is obtained when precision and recall reach the corresponding maximum values of 1 each.

The F-measure is used for evaluation as follows : first, an operating point on the ROC (say, equal

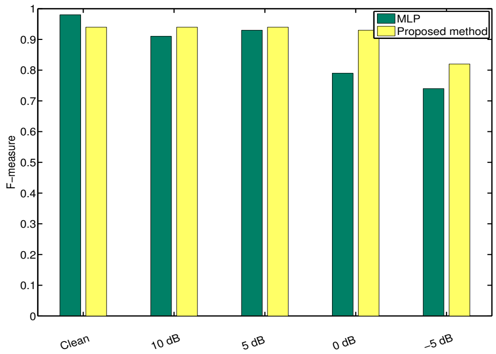

FIG. 5 – *Comparison of proposed and MLP based system using F-measure*

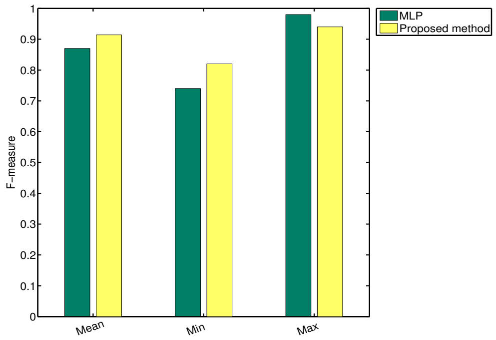

FIG. 6 – *Comparison of proposed and MLP based system using mean, minimum and maximum F-measure values over the entire data set.*

error rate - EER) is chosen. The threshold is determined for all the SND methods on the clean speech at this operating point. For all SNR levels in the test data set, the SND algorithms are deployed with these thresholds. The metrics, true positives, false negatives and false positives, are measured. The F-measure is obtained from these quantities.

The F-measure based comparison between the proposed method and the MLP based system is shown in Fig. 5. Here the operating point was EER on clean speech. It can be seen from the figure that while the F-measure in clean speech of the MLP based method is high ($EER = 2\%$) in comparison with the proposed method ($EER = 6\%$) , its performance at lower SNR levels drops below the proposed method. It indicates that the proposed method is less sensitive to thresholds than the MLP based method.

Fig. 6 summarizes the key F-measure based statistics. Here the mean, minimum and the maximum values of the F-measures of the MLP based and the proposed methods are computed over the following SNR conditions : Clean, 10 dB, 5 dB, 0 dB and -5 dB. It shows that the mean F-measure of the proposed method at all SNR levels is higher than that of the MLP based method's by about 5% relative (4.4%absolute). Further, it also shows that the F-measure of the MLP based method drops by 24% (absolute) from clean speech to -5 dB SNR. In comparison, the proposed method drops only by 12% for the same change in environment, again indicating the robustness of the thresholds.

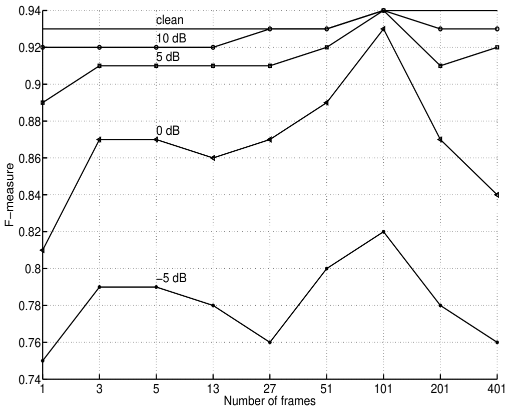

Fig. 7 – *Determination of optimal lengths at 5 different SNR levels : clean, 10 dB, 5 dB, 0 dB, -5 dB*

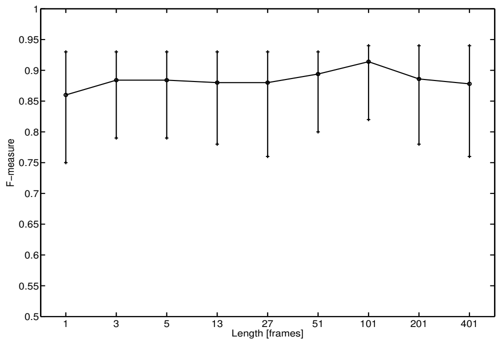

Fig. 8 – *Comparison of the mean, minimum and maximum values of F-measure for different lengths over the entire data set.*

4.5 Determination of optimal length

We now wish to determine the optimal length of the feature vector. This is done by varying the lengths of the contextual information at a particular SNR. Again the performance is measured using F-measure. The lengths of the feature vector (number of dimensions = number of full-band energy frames) studied in terms of number of frames of context were : 1 (energy threshold, with no context), 3, 5, 13, 27, 51, 101, 201 and 401. It should be noted that each frame is shifted by 10 ms.

Fig. 7 show the F-measure plots for various contextual lengths at three different SNR levels. From this plot, it can be observed that, for clean speech, using longer temporal context does not improve the performance. On the other hand, as the noise level increases, the temporal context becomes important. Further, it can be observed that the optimal context is around 101 frames (1000 ms).

As before, we illustrate the mean, maximum and minimum of the F-measure over all SNR levels for different lengths in Fig. 8. The figure shows that taking a 1000 ms context provides the most robust algorithm. Further, it can be seen that the simple energy based method is the most sensitive algorithm to changes in noise level and that when the context is increased to 1000 ms, the performance increases by 6.28% relative (5.4% absolute). Indeed, this result is not surprising.

From figures 6 and 8, it can be seen that the mean MLP method performance is better than the mean simple energy based method. Further, in clean environment, the MLP method outperforms the simple energy based method. On the other hand, at -5 dB SNR, MLP method performs worse than the energy based method, as the training and the testing conditions are badly mismatched.

5 Conclusion

We have presented a method for SND that employs full-band energy with long contextual information. This method utilizes LDA to obtain the weights of the context. The proposed method is compared with a state-of-the-art MLP based SND and an energy based system. In terms of F-measure, it shows an absolute performance gain of 4.4% and 5.4% respectively over these methods. It shows that even a simple feature such as full-band energy, when utilized with a large-enough context, is promising. In future work, we wish to investigate the importance of contextual information for sub-band energies.

6 Acknowledgements

This work was supported by the Dollbrain (NSF Project Micropower integrated face and voice detection, grant number : 200021-112354/1) and Detection and Identification of Rare Audio-Visual Cues (contract numbers of DIRAC is : FP6-0027787) projects ; managed by the IDIAP Research Institute on behalf of Swiss Federal Authorities.

Références

[1] J. Dines, J. Vepa, and T. Hain. The segmentation of multi-channel meeting recordings for automatic speech recognition. In *Int. Conf. on Spoken Language Processing (Interspeech ICSLP)*, pages 1213–1216, Pittsburgh, USA, 2006.

[2] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2001.

[3] J. S. Garofolo, C. D. Laprun, M. Michel, V.M. Stanford, and E. Tabassi. In *The NIST meeting room pilot corpus*, 2004.

[4] H. Hermansky and N. Morgan. RASTA processing of speech. In *IEEE Transactions on Speech and Audio Processing*, volume 2, pages 587–589, October 1994.

[5] H. Hermansky and S. Sharma. Traps - classifiers of temporal patterns. In *Int. Conf. on Spoken Language Processing (Interspeech ICSLP)*, Sydney, Australia, 1998.

[6] H. K. Maganti, P. Motlicek, and D. G. Perez. Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[7] N.Mesgarani, M. Slaney, and S.A. Shamma. Discrimination of speech from nonspeech based on multiscale spectro-temporal Modulations. In *IEEE Transactions on Audio, Speech and Language Processing*, volume 14, pages 920–930, May 2006.

[8] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones. The noisex-92 study on the effect of additive noise on automatic speech recognition. *Tech. Report DRA Speech Research Unit*, 1992.