Exploiting contextual information for speech/non-speech detection

Sree Hari Krishnan Parthasarathi, Petr Motlicek, and Hynek Hermansky

IDIAP Research Institute, Martigny
Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland
{hari.parthasarathi,petr.motlicek,hynek}@idiap.ch

Abstract. 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. The optimal contextual length was found to be 1000 ms. Further numerical optimizations yield an improvement (3.37% absolute), resulting in an absolute gain of 7.77% and 8.77% over the MLP based and energy based methods respectively. ROC based performance evaluation also reveals promising performance for the proposed method, particularly in low SNR conditions.

Key words: Speech/non-speech detection, modulation spectrum, temporal context

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. Historically, short-term energy has been one of the most important features for SND [1]. In this paper, we study the effect of long temporal context on signal energy for SND using a data-driven approach. Two recent studies of SND, [6] and [5], exploit temporal context using modulation spectrum on multiple spectral bands.

In our approach, the weights of the context around the frame-to-be-classified, are obtained using Linear Discriminant Analysis (LDA). This method gives us an interpretation in terms of a filter in the modulation spectral domain. However, it is well-known that when the features are correlated and when the dimension is large, the LDA covariance matrices are not estimated well. Many solutions to this problem of regularizing the covariance matrix exist [3]. The details of our regularization method are provided in Section 2.

The rest of the paper is organized as follows. Section 2 discusses the proposed method, with and without regularization, in detail. Description of the experimental evaluation and the data set is provided in Section 3. Finally, we draw some conclusions in Section 4.

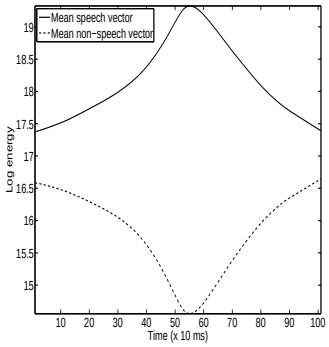

Fig. 1. *Mean speech and non-speech vectors.*

2 Obtaining the weights of the temporal context: Proposed method

2.1 Features

The first step is to obtain feature vectors 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 (obtained from the training data set, Section 3) 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.

2.2 Training 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, using LDA [3]. The label of the class at the center of the feature vector determines the training targets. LDA is used to obtain the weight vector for classification. This is the weight vector obtained without regularization.

However, since the features are highly correlated and are in a high dimension, the estimate of the within-class covariance matrices becomes poor. One of the solutions to this problem is to first perform dimensionality reduction using

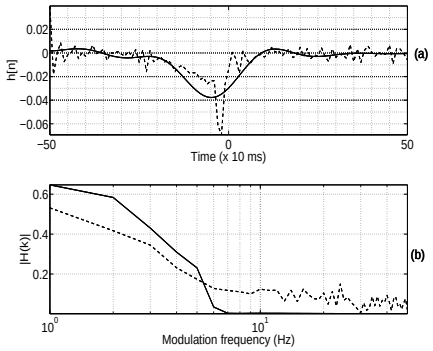

Fig. 2. *(a) Dotted and solid lines indicate impulse responses obtained by LDA and regularized LDA respectively (b) Dotted and solid lines indicate magnitude responses obtained by LDA and regularized LDA respectively*

Principal Component Analysis (PCA) and then employ LDA [3]. On the other-hand, it is well known that PCA is asymptotically equivalent to Discrete Cosine Transform (DCT) for Markov-1 signals if the correlation coefficient is close to 1[1]. Therefore, we perform dimensionality reduction by projecting the features obtained using the method described in Section 2.1 onto the first few DCT bases. LDA is then used to estimate the contextual weights in this space. The weights in the subspace are projected back on the entire DCT bases to estimate the weight vector in the original space. This is the weight vector obtained with regularization.

For comparison, Fig. 2(a) shows the weight vectors (flipped left to right, for interpretation as an impulse response) obtained by LDA and regularized LDA. It can be observed that regularizing the LDA using DCT yields a smooth weight vector. Further, it is also shown experimentally that the overall performance increases when regularized LDA is used (Section 3).

We now analyze the shape of the impulse response: 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 300 to 400 ms around the center is important for classification. Also, note that reducing the feature vector dimension to one, reduces the method to energy thresholding.

Determination of SND boundaries During testing, the feature vectors ($\mathbf{x}_i$) of the speech signal are computed as described in Section 2.1. The vectors are projected on to $\mathbf{w}$. These projected values are then compared with the threshold (θ), to determine the class.

[1] Our experiments also showed virtually identical weight vectors obtained by LDA after either PCA or DCT.

3 Experiments and evaluation

We investigate three questions in this section: (a) How does the proposed method compare with the state-of-the-art? Section 3.2 and 3.3 address this question. (b) Does utilizing context yield better performance? To answer this, the significance of the contextual window is studied in Section 3.4. (c) Does regularization improve classification performance? Towards this, regularized LDA is compared with LDA without regularization (Section 3.5) for the optimal length obtained in Section 3.4.

3.1 Experimental setup

Experiments are conducted on a subset of the NIST meeting room corpus [4]. Data obtained from close-talking microphones are 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 are obtained by forced-alignment of ASR phoneme models [2]. All phonemes except 'sil' are considered 'speech', while the 'sil' regions are 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 [7] is added at various SNR levels.

To evaluate the importance of temporal context for SND, we compare the proposed method (long temporal context) with two short-term methods (a) a state-of-the-art multi-layer perceptron (MLP) based method [2] 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.

3.2 Comparison with MLP based system using ROC curves

In the first set of experiments, a context of 1000 ms is used for the proposed method (without regularization). A comparison of the proposed method with the MLP based system using the Receiver Operating Characteristics (ROC) curve method is done. To plot the ROC curve, "true speech positives" and "false speech positives" are computed 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 Fig.3. It illustrates that the proposed system performs better than the MLP

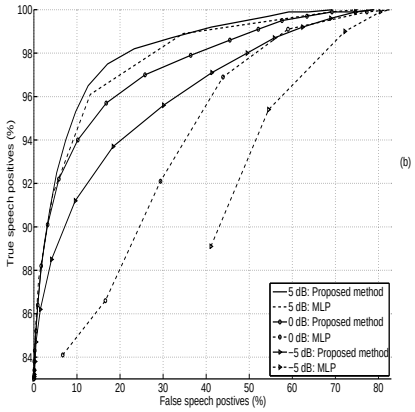

Fig. 3. *Comparison of proposed and MLP based system in 5, 0 and -5 dB SNR.*

based system in significantly noisy conditions. We attribute the performance of the proposed method to the usage of long-term contextual information.

Experiments also revealed 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 training conditions.

3.3 Comparison with MLP based system using F-measure

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 this aspect of the SND algorithms, F-measure is utilized. It is defined as the harmonic mean of "precision" and "recall". 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 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.

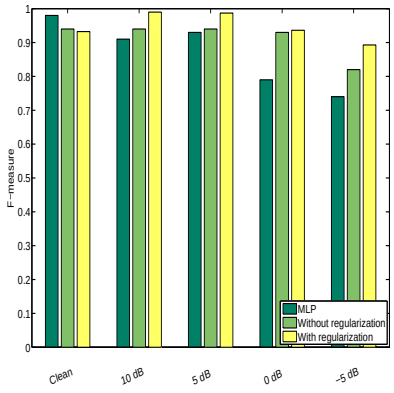

Fig. 4. *Comparison of MLP based system and the proposed method (with and without regularization) using F-measure*

The F-measure based comparison between the proposed method (with and without regularization) and the MLP based system is shown in Fig. 4. Here the operating point was EER on clean speech. In this section, we discuss the comparison of proposed method without regularization and the MLP based system. Section 3.5 discusses this graph with respect to regularization. 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.

Computation of the mean F-measure over different SNRs (Clean, 10 dB, 5 dB, 0 dB and -5 dB), shows that the mean F-measure of the proposed method(0.914) at all SNR levels is higher than that of the MLP based method(0.87) by about 4.4% absolute performance. Further, we observe 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.

3.4 Determination of optimal length

In the second set of experiments, the length of the context is varied to identify the optimal length for the proposed method. Studying the effect of different contextual lengths inherently includes a comparison with the short-term energy based method as well.

The optimal length is obtained 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

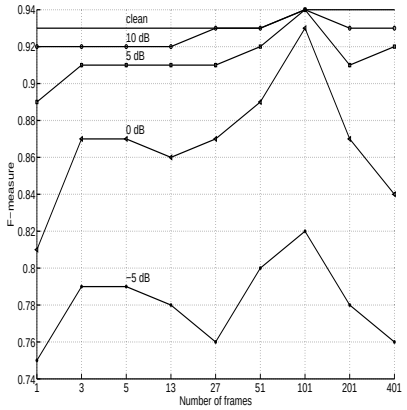

Fig. 5. *Determination of optimal lengths at 5 different SNR levels: clean, 10 dB, 5 dB, 0 dB, -5 dB*

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.

Fig. 5 shows the F-measure plots for various contextual lengths at five 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). Also, 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 5.4% absolute. Indeed, this result is not surprising.

Also, the performance of the mean MLP method is better than that of 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.

3.5 Experiments with regularized LDA

Experiments with the optimal weighted context (101 frames) derived using regularized LDA is discussed. The results of experiments, with and without regularization, is shown in Fig. 4. This figure shows that, except in the case of clean speech, regularization improves the performance. Computation of the mean F-measure over all the SNR conditions for the two cases are: (a) 94.77% with regularization (b) 91.4% without regularization. It shows that regularized LDA improves the performance by 3.37% absolute over LDA without regularization.

The overall improvement obtained by regularized LDA over MLP based system is 7.77% (absolute). Further, the regularized LDA yields an absolute improvement of 8.77% over the energy based method.

4 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. Further,to improve the estimation of the within class covariance matrices in LDA, regularization using DCT is performed. This modification yields an absolute improvement of 7.77% and 8.77% over the MLP based and energy based method respectively.

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.

5 Acknowledgements

This work was supported by the MIFAVO (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.

References

1. B.S. Atal and L.R. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Trans. on Acoust., Speech and Signal Process.*, 1976.
2. 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.
3. K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.
4. J. S. Garofolo, C. D. Laprun, M. Michel, V.M. Stanford, and E. Tabassi. In *The NIST meeting room pilot corpus*, 2004.
5. 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.
6. N.Mesgarani, M. Slaney, and S.A. Shamma. Discrimination of speech from non-speech based on multiscale spectro-temporal Modulations. In *IEEE Transactions on Audio, Speech and Language Processing*, volume 14, pages 920–930, May 2006.
7. 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.