IDIAP RESEARCH REPORT

A Data-driven Approach to Speech/Non-speech Detection

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

IDIAP–RR 08-23

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 08-23

A Data-driven Approach to Speech/Non-speech Detection

Sree Hari Krishnan Parthasarathi,
Petr Motlicek, and
Hynek Hermansky

April 2008

Résumé. We present a data-driven approach to weighting the temporal context of signal energy to be used in a simple speech/non-speech detector (SND). The optimal weights are obtained using linear discriminant analysis (LDA). Regularization is performed to handle numerical issues inherent to the usage of correlated features. The discriminant so obtained is interpreted as a filter in the modulation spectral domain. Experimental evaluations on the test data set, in terms of average frame-level error rate over different SNR levels, show that the proposed method yields an absolute performance gain of 10.9%, 17.5%, 7.9% and 8.3% over ITU's G.729B, ETSI's AMR1, AMR2 and a state-of-the-art multi-layer perceptron based system, respectively. This shows that even a simple feature such as full-band energy, when employed with a large-enough context, shows promise for applications.

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, [2] and [3], 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.

The rest of the paper is organized as follows. In Section 2, a brief overview of LDA is provided. Experimental data set is described in Section 3. Section 4 discusses the proposed method in detail. Description of the experimental evaluation is provided in Section 5. Finally, we draw some conclusions in Section 6.

2 Review of linear discriminant analysis

LDA [4] 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 following discussion, an overview of LDA is provided for a two-class problem.

Let $\left\{{\mathbf{x}_i}^k\right\}$ 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 [4], we define the within-class (s_w) and between-class (s_b) scatter matrices as follows :

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

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 $J(.)$:

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

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

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

3 Experimental data set

The data used in our experiments were a subset of the NIST meeting room corpus [5]. Speech obtained from close-talking microphones, sampled at 16 kHz, were used. 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

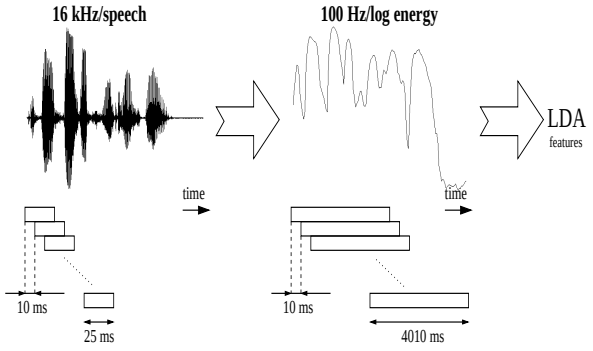

FIG. 1 – *Feature extraction procedure.*

of ASR phoneme models [6]. This procedure yields consistent labels. All phonemes except 'sil' were considered '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] was added at various SNR levels : 10 dB, 5 dB, 0 dB, -5 dB. The combined clean and noisy training and test data set is thus 5 and 15 hours, respectively.

4 Obtaining the weights of the temporal context : Proposed method

4.1 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 using overlapped windows (i.e., size of 4010 ms and a shift of 10 ms) on this temporal trajectory. This feature extraction method introduces a context around the frame under consideration. We utilize a context of 2000 ms in the past and in the future, resulting in a resolution of 0.25 Hz in the modulation spectral domain [8]. The dimension of this feature vector is thus 401 : for a frame at the center, there are 200 frames (x 10 ms) on either side. These feature vectors are used as parameters for the LDA classifier. The feature extraction procedure is illustrated in Fig. 1.

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. 2. These vectors are 4010 ms long. 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. Moreover, it can be seen that the effect of the class (speech/non-speech) is less at distances more than 1000 ms from the center.

4.2 Training LDA to obtain the weights of the context

In this section, we obtain the weights of the context around the frame that is to be classified. LDA is used to obtain the weight vector (or discriminant) for classification. The LDA discriminant is derived from the training data (Section 3). The label of the class at the center of the feature vector determines the training targets.

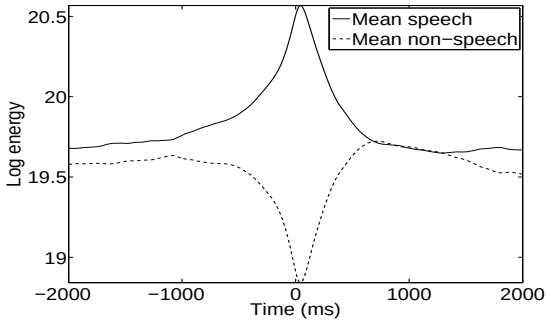

FIG. 2 – *Mean speech and non-speech vectors.*

4.2.1 Regularization

We note that the discriminant so obtained is very noisy. This is due to the dimensions of the features being highly correlated. It causes the estimate of the within-class covariance matrices to become badly conditioned. One of the solutions to the problem of correlated features is to first perform dimensionality reduction using Principal Component Analysis (PCA) and then employ LDA [4]. 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 feature vectors on to the first few DCT bases. In our case, the number of DCT bases used is 100. A dimension reduction from a space of 401 dimensions to a subspace of 100 is performed. LDA is then used to estimate the weights in this subspace. The weights in the subspace are projected back on to entire DCT bases to estimate the weight vector in the original space. This is the weight vector obtained with regularization. We now present two successive training set-ups :

– Training regularized LDA on clean environment, denoted as "LDA-clean".
– Train regularized LDA on noisy data, denoted as "LDA-noisy".

Both the trained LDA models are used for comparison with standard and the MLP based SNDs.

4.2.2 Training on clean speech : "LDA-clean"

As a first step, LDA is trained on features obtained on 1 hour of clean data (Section 3). The discriminant represents an impulse response of a filter on the log energy signal, sampled at 100 Hz. Note that from equation 3, reducing the feature vector dimension to one, simplifies the method to energy thresholding.

The impulse and the magnitude frequency responses of the discriminant are plotted in Fig. 3. The valley at the center of the impulse response suggests that a context of 600 ms around the center is important for classification. Further, the discriminant indicates that in clean environments the DC component is important for SND.

4.2.3 Training on noisy speech : "LDA-noisy"

The same clean data at different noise levels (clean, 10 dB, 5 dB, 0 dB and -5 dB), equal to 5 hours, is supplied as training examples to LDA. The feature extraction procedure is identical to the previous set-up. This forces LDA to learn an invariance to different noise levels. The discriminant is plotted in Fig. 4. The magnitude frequency response shows that the DC levels are not used for the discrimination. The impulse response suggests a positive weighting around center (600 ms) and a negative weighting outside this region.

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

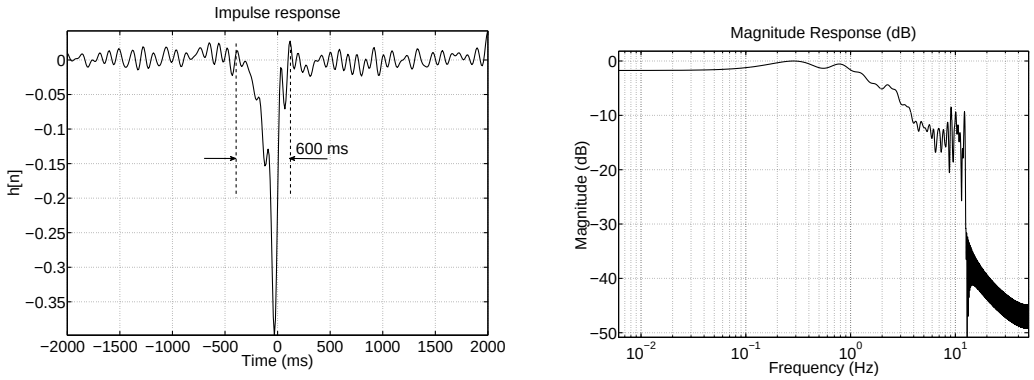

FIG. 3 – *Impulse and magnitude frequency responses of the regularized discriminant on clean speech.*

4.2.4 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 feature vectors.
3. θ is determined as the point where the two distributions cross each other.

4.2.5 Determination of SND boundaries

During testing, the feature vectors ($\mathbf{x}_i$) of the speech signal are computed as described in Section 4.1. The vectors are projected on to $\mathbf{w}$. These projected values are then compared with the threshold (θ), to determine the class.

5 Experiments and Evaluation

5.1 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 three short term speech/non-speech detection methods :

1. SND from ITU's G.729B codec [9] uses a piecewise linear discriminant based on line spectral frequencies, high and low band energies and zero-crossing rate. It makes decisions every 10 ms on speech sampled at 8 kHz.
2. SND modules of ETSI's AMR1 and AMR2 codec [10] use sub-band energies and an adaptive threshold to make SND decisions. These procedures make decisions every 20 ms on speech sampled at 8 kHz.
3. A state-of-the-art multi-layer perceptron (MLP) system [6] uses 12 MF-PLP coefficients along with their first and second derivatives. In addition, the following auxiliary features are added : normalized energies from all channels, signal kurtosis, mean cross-correlation and maximum normalized cross-correlation. The MLP is trained on approximately 98 hours of training data and makes decisions every 300 ms on speech sampled at 16 kHz.

The data described in Section 3 is downsampled to 8 kHz for use with G729B, AMR1 and AMR2.

5.2 Evaluation

Since the proposed ("LDA-clean" and "LDA-noisy") and the MLP based methods are threshold based, we do not use a receiver operating characteristic (ROC) curve based evaluation. The ROC based evaluation does not reveal the sensitivity of the threshold. More particularly, the threshold is set at a

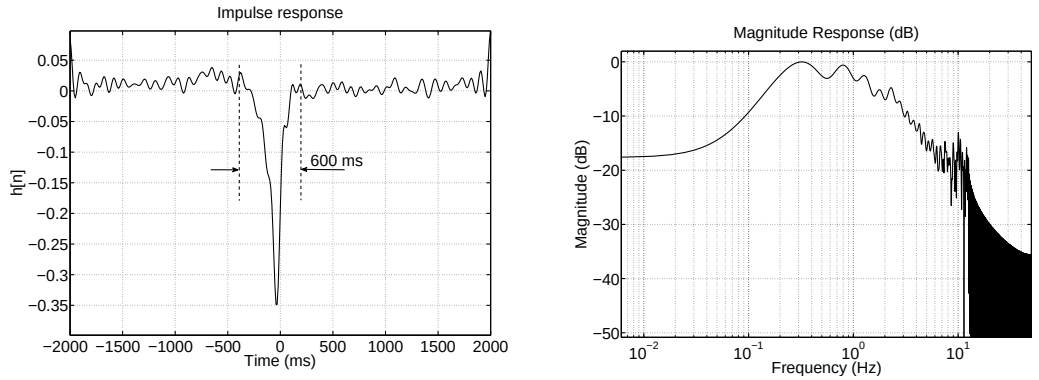

FIG. 4 – *Impulse and magnitude frequency responses of the regularized discriminant on noisy speech.*

particular operating point in the ROC curve, on the development data. When the testing environment is different from the development environment, the threshold changes. However, evaluations using ROC curves for the testing data involves varying the thresholds of the algorithm, thus obscuring what the threshold yielded.

In contrast to ROC based evaluation, we utilize a Frame-level Error Rate (FER) as a figure of merit to compare the performance of the algorithms :

$$\text{FER} = \frac{\text{False positives} + \text{False negatives}}{\text{Total number of frames}} \qquad [\%]. \tag{4}$$

The usage of FER is justified because the distribution of the speech and non-speech data is balanced. This metric is utilized as follows : first, an operating point on the ROC curve (say, equal error rate - EER) is chosen. The thresholds at EER are determined for both the proposed and the MLP based methods on the respective ROC curves for clean speech. For all SNR levels in the test data set, the proposed and the MLP systems are deployed with this threshold. The metrics (true positives, true negatives, false negatives and false positives) are measured. FER is computed. It is noted that FER of 50% is attained by an SND that (a) makes uniformly random speech or non-speech decisions ; (b) always declares speech ; (c) always declares non-speech. This is because our data set is balanced, i.e., speech and non-speech are approximately equal.

The comparison of the SND methods on the total test data set (15 hours, Section 3) is shown in Fig. 5, where "LDA-clean" and "LDA-noisy" refer to the regularized LDA trained on clean and noisy data, respectively. The figure shows that the MLP based system performs the best in the clean environment (2.5%). However, a machine learning technique would perform poorly when the testing conditions do not match the training conditions. This is consistent with our results which show that the MLP system performs the worst at low SNR levels.

Further it is not surprising that in clean environment "LDA-clean" yields a better performance (6.9%) than the "LDA-noisy" (8.2%). But "LDA-clean" performs poorly in noisy conditions (at -5 dB SNR, 32.8%). This is expected as the "LDA-clean" discriminant is sensitive to the DC levels. On the other hand, in "LDA-noisy", we provide the knowledge to LDA about the invariance to DC levels in addition to the modulation spectral characteristics of speech and noise. This improves its performance over other methods in more noisy conditions.

Fig. 6 summarizes the key FER based statistics. Here the mean, minimum and the maximum values of the FER of the SND methods are computed over different SNR conditions (Clean, 10 dB, 5 dB, 0 dB and -5 dB). It shows that the mean FER over all SNR levels of the "LDA-noisy" (8.6%), is clearly lower than the best standard method (AMR2 with a mean FER of 16.5%). However, this comes at the cost of increased algorithmic latency (2000 ms) for the proposed method.

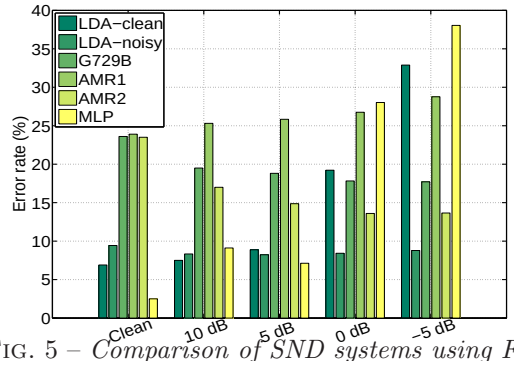

FIG. 5 – *Comparison of SND systems using FER.*

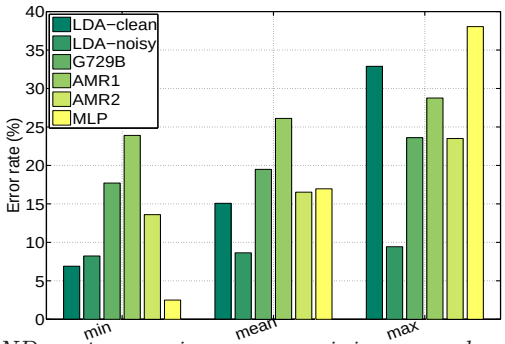

FIG. 6 – *Comparison of SND systems using mean, minimum and maximum FER values over the entire data set.*

6 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. To test the effectiveness of contextual information, we compare the proposed method with two short-term SND standards (ITU's G.729B and ETSI's AMR1 and AMR2) utilizing ~ 20 ms of temporal context, and a state-of-the-art MLP based method that utilizes 300 ms of temporal context.

In terms of frame-level error rate, experimental evaluation shows that the proposed method yields an absolute performance gain of 10.9%, 17.5%, 7.9% and 8.3% over G.729B, AMR1, AMR2 and MLP based systems, respectively. It shows that even a simple feature such as full-band energy, when utilized with a large-enough context, is promising. However, this comes at the cost of algorithmic delay (2000 ms).

In future work, we wish to investigate the importance of contextual information for sub-band energies.

7 Acknowledgements

This work was supported by the MIFAVO (NSF - Micropower integrated face and voice detection, grant number : 200021-112354/1) ; and DIRAC (Detection and Identification of Rare Audio-Visual Cues, contract number : FP6-0027787) projects. The projects are managed by the IDIAP Research Institute.

Références

[1] B. Atal and L. 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] N.Mesgarani, M. Slaney, and S. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal Modulations," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, May 2006, pp. 920–930.

[3] 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, pp. 1037–1040.

[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

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

[6] 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)*, Pittsburgh, USA, 2006, pp. 1213–1216.

[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.

[8] H. Hermansky, "The modulation spectrum in the automatic recognition of speech," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.

[9] A. Benyassine, E. Shlomot, and H. Su, "ITU Recommendation G.729 Annex B : A Silence Compression Scheme for Use with G.729 Optimized for V.70 Digital Simultaneous Voice and Data Applications," *IEEE Comm. Mag.*, pp. 64–73, 1997.

[10] "Digital Cellular Telecommunications System (Phase 2+) ; Voice Activity Detector (VAD) for Adaptive Multi Rate (AMR) Speech Traffic Channels," 1999.