IDIAP RESEARCH REPORT

Unsupervised Speech/Non-speech Detection for Automatic Speech Recognition in Meeting Rooms

Hari Krishna Maganti [1,2] Petr Motlicek [1]
Daniel Gatica-Perez [1]

IDIAP–RR 06-57

September, 2006

1 IDIAP Research Institute
2 University of Ulm, Ulm, Germany

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 06-57

Unsupervised Speech/Non-speech Detection for Automatic Speech Recognition in Meeting Rooms

Hari Krishna Maganti   Petr Motlicek   Daniel Gatica-Perez

September, 2006

Abstract. The goal of this work is to provide robust and accurate speech detection for automatic speech recognition (ASR) in meeting room settings. The solution is based on computing long-term modulation spectrum, and examining specific frequency range for dominant speech components to classify speech and non-speech signals for a given audio signal. Manually segmented speech segments, short-term energy, short-term energy and zero-crossing based segmentation techniques, and a recently proposed Multi Layer Perceptron (MLP) classifier system are tested for comparison purposes. Speech recognition evaluations of the segmentation methods are performed on a standard database and tested in conditions where the signal-to-noise ratio (SNR) varies considerably, as in the cases of close-talking headset, lapel, distant microphone array output, and distant microphone. The results reveal that the proposed method is more reliable and less sensitive to mode of signal acquisition and unforeseen conditions.

1 Introduction

Aiming at discriminating speech and non-speech segments from a given audio signal, speech/non-speech detection (SND) is crucial for speech signal processing applications. Inaccurate boundaries are an important cause of errors in automatic speech recognition systems, and a pre-processing stage that segments the signal into periods of speech and non-speech is invaluable in improving the recognition accuracy. An evaluation of an isolated-word recognizer has shown that more than half of the recognition errors are due to inaccurate word boundaries [1]. Apart from ASR, a good segmentation of audio stream has many practical applications such as broadcast news transcription [2], automatic audio indexing and summarization [3], audio and speaker diarization [4]. Accordingly, segmentation has to be easily integrated into the systems concerned, but it should not increase the overall computational load.

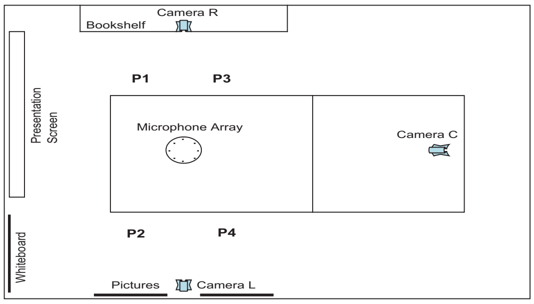
One of the issues in the design of a SND system is the selection of an appropriate feature set that captures the temporal and spectral structure of the signals. Scheirer and Slaney investigated features for speech/music discrimination that are closely related to the nature of human speech [5]. The proposed features, including, spectral centroid, spectral flux, zero-crossing rate, 4Hz modulation energy (related to the syllable rate of speech), and the percentage of low-energy frames, have been explored in the task of discriminating speech from various types of music. In [6], entropy and dynamism features based on posterior probabilities of speech phonetic classes (as obtained at the output of an HMM/ANN large vocabulary continuous ASR system) are used to form an observation vector sequence, which is used in a HMM classification framework. Depending on the process involved, the SND techniques can be divided into two general categories: threshold detection process, and pattern-recognition process. In the threshold detection process, the acoustic features for each frame of speech signal are extracted and then compared with preset thresholds to classify each frame. The frame feature parameters include energy, zero-crossings, pitch, entropy, duration, and linear prediction error energy [7, 5]. In the pattern-recognition process, the estimates of model parameters for speech and non-speech are required. The most commonly used features for discriminating speech from music, and other sound sources are the cepstrum coefficients such as Mel-frequency cepstral coefficients (MFCCs) and perceptual linear prediction (PLPs) cepstral coefficients, which are extensively used in speaker- and speech recognition tasks. Although these signal representations have been originally designed to model the short-term spectral information of speech events, they were also successfully applied in SND systems in combination with Gaussian Mixture Models (GMMs) or Hidden Markov Models (HMMs) for separating different sound sources (broadband speech, telephone speech, music, noise, silence, etc.) [2, 6]. In the context of conference rooms, combination of energy features generated directly from the signal, and the acoustic phonetic features derived from observations generated by ASR acoustic models were used as input to the GMM classification framework [8].

The existing methods are limited by two common drawbacks. On one hand, threshold based detection techniques fail under low SNR conditions, and on the other hand, pattern-matching techniques require large training data to train the models and need a prior knowledge of the noise. In this paper, a simple, robust, and accurate algorithm based on modulation spectrum for SND tasks is proposed, which performs well in low SNR conditions and neither requires training data nor a prior knowledge of the noise. The special characteristic of long-term modulation spectrum, that speech segment is dominated by components between 2 and 16 Hz, which reflect syllabic and phonetic temporal structure of speech is used [9]. This approach reduces the computational complexity and time as required for pattern matching methods. The performance is also close to real-time, as the decision is made on short-segments of the signal (200 - 1000 ms), rather than over the entire utterance, which is a basic requirement for conventional threshold detection methods [7, 5].

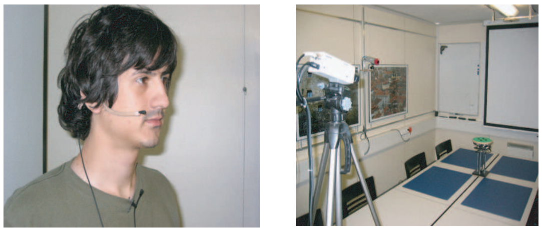
The paper is organized in four sections: In Section 2, an overview of the system setup for the database used for evaluations is given. Section 3 explains the algorithm design. Section 4 presents experiments and results. Finally, Section 5 concludes the paper.

2 System setup

The data used for experiments is recorded in an instrumented meeting room comprising of a microphone array, and headset and lapel microphones. All the microphones are of high quality electret type. The sensor configuration is similar to the system presented in [10]. Figure 1.a shows the layout, the positions of the microphone array, and the typical speaker positions in the meeting room.



a



b

Figure 1: *Schematic diagram of the meeting room. The headset microphone is close to the mouth of the speaker, the lapel microphone fixed at the collar of the speaker is about 15 - 20 cm away from the mouth, and the distant microphone and microphone array are about 90 - 100 cm away from the speaker.*

3 Algorithm Design

The proposed approach is based on long-term modulations, examining the slow temporal evolution of the speech energy with time-windows in the range of 200 - 800 ms, contrary to the conventional short-term modulations (frequently used in ASR) studied with time-windows up to 10 - 30 ms which capture rapid changes of the speech signals. The relative prominence of slow temporal modulations is different at various frequencies, similar to perceptual ability of human auditory system. Particularly, most of the useful linguistic information is in the modulation frequency components from the range between 2 and 16 Hz, with dominant component at around 4 Hz [11, 12, 13]. In [12], it has been shown that for some realistic environments, the use of components from the range below 2 or above 16 Hz can degrade the recognition accuracy. The proposed algorithm is based on this particular characteristic of speech, which is used to classify speech and non-speech signals in order to characterize each acoustic event. The block diagram of the algorithm is shown in Figure 2, and is as follows:

For a given 16 kHz sampled signal $x(t)$, the Fast Fourier Transform (FFT) is computed over N points and the segment is shifted by n ms, resulting in a $\frac{N}{2}$ dimensional FFT vector. The Mel-scale transformation is applied to the FFT vector. The logarithmic-like Mel scale models the non-linear frequency resolution of the human ear,which is defined by [14]

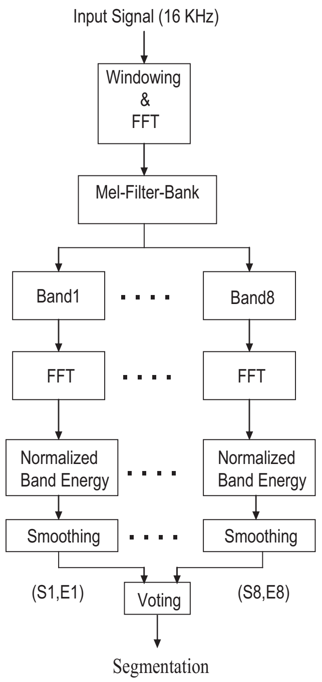


Figure 2: *Block diagram of the speech/non-speech detection algorithm.*

$$\text{Mel}(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{1}$$

The filters used in Mel-frequency analysis are generally triangular in shape, and are equally spaced along the Mel-scale. The output is a Mel-scaled vector consisting of K bands. The computations are made over all the incoming signal, resulting in a sequence of energy magnitudes for each band sampled at $\frac{1}{n}$ Hz. In each band, the modulations of the signal are analyzed by computing FFT over the P points and the segment is shifted by p ms. The result is a sequence of $\frac{P}{2}$ dimensional modulation vectors. The energies for the frequencies between the 2 - 16 Hz represent important components for the speech signal. An example of the modulation spectrum of a audio signal for values K=1, N = 512, P = 100 is shown in Figure 3.b. It can be observed that the speech and non-speech segments are clearly distinguished by high and low activities in the regions which correspond to 2 - 16 Hz.

The modulation energy corresponding to 2 - 16 Hz is computed and normalized with the total modulation energy (1 - 50 Hz), resulting in a time sequence $s(t)$ (sampled at $\frac{1}{n}$ Hz). This sequence is smoothed using the moving average method with span of $2P$ to give $\hat{s}(t)$, as shown in Figure 3.c.

The mean T of the $\hat{s}(t)$ is estimated from few utterances of the headset development data, which is used for the decision $D(t)$, as defined by

$$D(t) = \begin{cases} 0 & \hat{s}(t) < T \\ 1 & \hat{s}(t) \geq T \end{cases} \tag{2}$$

where $\hat{s}(t)$ is the smoothed version of the normalized energy. This computation is done for each of the K bands. Finally, a majority voting is performed (every n ms) to decide the segmentation boundaries.

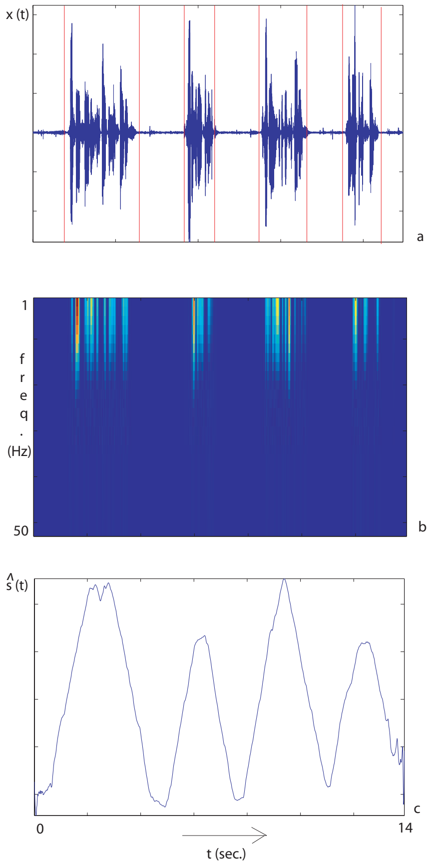


Figure 3: *(a) Waveform of the given audio signal, and red lines indicate segmentation boundaries given by D(t) (b) Log modulation spectrum of speech utterance for K = 1, N = 512, and P = 100. The frequency range between 2 - 16 Hz illustrate high activity for speech segments (c) Smoothed normalized energy.*

4 Experiments and results

All the experiments were conducted on a subset of the Multi-Channel Wall Street Journal Audio-Visual (MC-WSJ-AV) corpus. The specification and structure of the full corpus are detailed in [15]. A part of *Single speaker stationary* data, in which the speaker reads out sentences from different positions within the meeting room is used. Most of the data comprised non-native English speakers with different speaking styles and accents. The data is divided into development (DEV) and evaluation (EVAL) sets with no common speakers in both the sets.

Short-term energy, short-term energy and zero-crossing based segmentation techniques [7], and a recently proposed Multi Layer Perceptron (MLP) based system [16] were evaluated to compare the efficiency of the proposed novel algorithm in detecting speech non-speech segments. Relying on a MLP classifier, this system was trained from several meeting room corpora to identify speech/non-speech segments. The training was performed with a corpora comprising headset recordings, which included approximately 112 hours of speech over 150 meetings.

With a view to evaluate the proposed method, the parameters as described in Section 3 are set as: $N = 512$, $n = 10$, $K = 8$, $P = 50$, p=10. These parameters were selected in order to handle a wide variety of modes of speech acquisition in the meeting room, right from the headset to distant microphone which was around 100 cm away from the speaker. To determine the optimum decision criteria D, the T which is the mean of the normalized modualtion energy was computed from 10 utterances of the development headset data. This was fixed for all the experiments performed irrespective of the testing channel i.e, lapel, distant microphone or output of the microphone array.

For the speech recognition experiments to evaluate the performance of the above mentioned techniques, a full HTK based recognition system [14], trained on the original Wall Street Journal database (WSJCAM0) was used. The training set consisted of 53 male and 39 female speakers, all with British English accents. The system consisted of approximately 11000 tied-state triphones with three emitting states per triphone and six mixture components per state. 52-element feature vectors were used, comprising of 13 MFCCs (including the 0th cepstral coefficient) with their first, second, and third order derivatives. Cepstral mean normalization is performed on all the channels. The dictionaries used are generated from that developed for the Augmented Multi-party Interaction (AMI) project and used in the evaluations of National Institute of Standards and Technology rich transcriptions (NIST RT05S) system [17], and the language models are the standard MIT-Lincoln Labs 5k and 20k Wall Street Journal trigram language models. To reduce the channel mismatch between the training and test conditions, the baseline HMM models were adapted using a maximum likelihood linear regression (MLLR) [18]. A static two-pass approach was used, where in the first pass a global transformation was performed and in the second pass a set of specific transforms for speech and silence models were calculated. This was followed by maximum-a-posteriori (MAP)[19] adaptation, where MLLR transformed means were used as the priors. For the experiments, 15 minutes of data of seven speakers was used for adaptation, and 10 minutes data of five speakers for testing.

Table 1: Speech recognition results. The values in the first column represent baseline WER, obtained from manual segmentation, and other values are compared with respect to these values.

	WER (%)				
Signal	**M**	**E**	**E+ZC**	**MLP**	**MS**
Headset	**21.3**	12.8	6.1	0.6	0.8
Lapel	**27.9**	11.4	4.8	1.8	0.6
Distant Microphone	**38.6**	8.0	5.3	7.2	0.4
Beamformer Output	**26.8**	6.5	4.1	2.4	0.3

Speech recognition experiments on different channels including headset, lapel, distant microphone and the output of the beamformer [10] were performed to evaluate the performance of the various techniques. The results are presented in Table 1, wherein M, E, E+ZC, MLP, MS represent manual, energy, energy + zero-crossing, multi layer perception, and modulation spectrum based segmentations, respectively. The values in the first column represent the baseline word error rates, which were obtained from the manual segmentation of the speech data. All other values were compared with reference to these values. From Table 1, it is clear that the energy based approach (method E) performed poorly for all the channels. Adding the zero-crossing feature to energy (method E+ZC) helped in reducing the WER by about 50 % in all the cases. The MLP based approach performed close to manual segmentation as it was trained on headset data of the large corpus [16]. However, the performance has decreased, when the same MLP (headset trained) was used for lapel, distance microphone, and microphone array output, for obvious reasons.

The results as presented in Table 1, demonstrate that the proposed modulation spectrum based approach was accurate and close to manual segmentation for all the channels. Nevertheless, training based approaches require large training data to model prior knowledge of the noise, and the process itself involves huge amounts of computational resources and prolonged duration, for reliable and robust performance. Other limitation of training based approach is that it is more sensitive to different training and unforeseen conditions. For example, in situations, where the speaker is using white board or moving around, the training based approaches cannot perform efficiently as the distance between the microphone and the speaker is not known exactly. Apart from accuracy and robustness, the proposed approach provides additional benefits. Requiring neither training data nor prior knowledge of the noise, it can be used for SND tasks of any unknown channel and unforeseen conditions. The performance is close to real-time as the classification decision is made on short segments of the signal, rather than over the complete utterance.

5 Conclusions

This paper has presented a novel algorithm based on modulation spectrum for speech/non-speech detection. This algorithm has been compared to manual segmentation, short-term energy, short-term energy and zero-crossing based segmentation techniques, and a recently proposed MLP classifier trained system. The speech recognition based evaluations are performed on real data in a meeting room for stationary speaker for all the methods and varying signal-to-noise ratios i.e., headset, lapel, distant microphone, and beamformer output. The results illustrate that the proposed simple technique is accurate, robust, close to real-time, and can be applied for SND tasks for any mode of speech acquisition and unforeseen conditions. Our study also raised a number of issues, including the approaches for decision without using the evaluation data (presently mean of the smoothed normalized energy is used), and the number of parameters involved in the method to suit different environments and acquisition channels. These subjects will be studied in the future.

Acknowledgment

This work was supported by the EC projects AMI, DIRAC, and the Swiss NCCR IM2. We thank John Dines for his help with the MLP system, Mike Lincoln and Iain McCowan for the collaboration in designing the MC-WSJ-AV corpus, and Bastien Crettol for his support to collect the data.

References

[1] J-C.Junqua, "Robustness and cooperative multimodal man-machine communication applications," SMMD, 1991, vol. I, pp. 101–112.

[2] P.Beyerlein, X. Aubert, and R.Haeb-Umbach, "Large vocabulary continuous speech recognition of broadcast news the philips/rwth approach," *Speech Communication*, vol. 37, pp. 109–131, 2002.

[3] J.Makhoul and et al, "Speech and language technologies for audio indexing and retrieval," IEEE, 2000, vol. 88, pp. 1338–1353.

[4] R.Sinha, S.E.Tranter, M.J.F.Gales, and P.C.Woodland, "The cambridge university march 2005 speaker diarisation system," Interspeech, 2005, pp. 2437–2440.

[5] E.Scheirer and M.Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," ICASSP, 1997, vol. 1, pp. 1331–1334.

[6] J.Ajmera, I.McCowan, and H.Bourlard, "Speech/music segmentation using entropy and dynamism features in a hmm classification framework," *Speech Communication*, vol. 40, pp. 351–363, May 2003.

[7] L.R.Rabiner and M.R.Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Systems Tech. Jour.*, vol. 54, pp. 297–315, February 1975.

[8] S.M.Chu, E.Marcheret, and G.Potamianos, "Automatic speech recognition and speech activity detection in the chil smart room," MLMI, 2005.

[9] S. Greenberg, "On the origins of speech intelligibility in the real world," ESCA-NATO Tutorial and Research Workshop on Robust speech Recognition for Unknown Communication Channels, 1997.

[10] I.McCowan, H.Maganti, D.Gatica-Perez, and et al, "Speech acquisition in meetings with an audio-visual sensor array," ICME, 2005.

[11] R.Drullman, J.Festen, and R.Plomp, "Effect of reducing slow temporal modulations on speech reception," *Journal Acoust. Soc.*, vol. 95, pp. 2670–2680, 1994.

[12] N.Kanedera, T.Arai, H.Hermansky, and M.Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communications*, vol. 28, pp. 43–55, 1999.

[13] H.Hermansky, "Auditory modeling in automatic recognition of speech," ECSAP, 1997.

[14] S.Young and et al, *The HTK Book Version 2.2*, Entropic Ltd, 1999.

[15] M.Lincoln, I.McCowan, J.Vepa, and H.Maganti, "The multi-channel wall street journal audio-visual corpus (mc-wsj-av): Specification and initial experiments," ASRU, 2005.

[16] J.Dines, J.Vepa, and T.Hain, "The segmentation of multi-channel meeting recordings for automatic speech recognition," ICSLP, 2006.

[17] T.Hain and et al, "The development of the ami system for the transcription of speech in meetings," MLMI, 2005.

[18] C.J.Leggetter and P.C.Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[19] J.-L.Gauvain and C.-H.Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 2, pp. 291–298, April 1994.