Detection of Speech Embedded in Real Acoustic Background Based on Amplitude Modulation Spectrogram Features

Jörn Anemüller, Denny Schmidt, Jörg-Hendrik Bach

Medical Physics Section, Carl von Ossietzky University Oldenburg, 26111 Oldenburg, Germany
joern.anemueller@uni-oldenburg.de, j.bach@uni-oldenburg.de

Abstract
A classification method is presented that detects the presence of speech embedded in a real acoustic background of non-speech sounds. Features used for classification are modulation components extracted by computation of the amplitude modulation spectrogram. Feature selection techniques and support vector classification are employed to identify modulation components most salient for the classification task and therefore considered as highly characteristic for speech. Results show that reliable detection of speech can be performed with less than 10 optimally selected modulation features, the most important ones are located in the modulation frequency range below 10 Hz. Detection of speech in a background of non-speech signals is performed with about 90% test-data accuracy at a signal-to-noise level of 0 dB. Compared to standard ITU G729.B voice activity detection, the proposed method results in increased true positive and reduced false positive rates induced by a real acoustic background.

Index Terms: Speech processing, speech detection, pattern classification, feature extraction, amplitude modulation, acoustic objects

1. Introduction

The human auditory system is remarkably sensitive in detecting the presence of specific acoustic objects even when they are embedded in real background sounds. As speech is arguably the most important acoustic object, the detection of its presence is of importance for a variety of applications such as noise reduction and speech recognition. A real-world acoustic background is itself composed of a multitude of interfering acoustic objects that may appear and disappear from the scene (such as cars in a road traffic scene) resulting in an often strongly fluctuating nature of the background. The present work proposes a method for detecting the presence of speech in real acoustic backgrounds. It is, however, not limited to speech as a target object and is used in ongoing work for uni- and multi-modal detection of other (non-speech) objects in backgrounds [1, 10]. Further, we investigate the question which components of speech are particularly salient for its detection and employ feature selection methods to this end. Finally, a comparison with a standard voice activity detection (VAD) algorithm demonstrates that the detection task in real fluctuating backgrounds is handled in a less than optimal way by the standard energy based approach which is outperformed by the proposed method in all investigated (i.e., low and high SNR) background conditions.

The novel contribution of the present work lies in (a) its systematic focus on identifying speech utterances in a real acoustic background of varying signal strength, (b) its specific adaptation of the amplitude modulation spectrogram feature extraction stage and (c) its data-driven analysis of which center-frequency and modulation-frequency components of the signal are most important for the task.

2. The amplitude modulation spectrogram

The choice of an appropriate signal representation is crucial to any classification method. Features chosen here are based on the amplitude modulation spectrogram ("AMS", [6, 9]) that has been adapted as outlined below in order to be largely invariant to speaker and channel variations. The choice of a modulation-based representation is motivated by the well-known importance in human and machine recognition of speech [4, 8] of modulation frequencies, \( f_m \), in the range of 2Hz up to 8Hz. Different modulation-based representations have been employed for several tasks in speech processing and acoustic scene analysis, cf. e.g. [2, 5, 11, 12].

The amplitude modulation spectrogram analyzes sound signals with respect to their modulation content by decomposing them into a 3-dimensional representation along the axes of time, frequency and modulation-frequency. The variant of the AMS employed here is tailored to be largely invariant and therefore robust with respect to (a) pitch variations between different speakers and (b) spectral distortions ("channel noise") of the signal. It is computed (cf. Fig. 1) by first extracting the spectral envelopes of the acoustic signal via an FFT (32 ms Hann window, 4 ms shift), squared magnitude and Bark-band computation [14]. A subsequent logarithmic compression transforms channel noise into an additive term in each spectral band. A second FFT (1 s Hann window, 500 ms shift) is applied within each spectral band to split it into different modulation spectral bands. Additive terms are isolated into the DC-band which is subsequently discarded, thereby making the representation largely invariant with respect to channel noise. The method finishes with an envelope extraction and log-compression step. A single slice of the amplitude modulation spectrogram captures the spectral and modulation spectral information within a one second long window. By sliding the window over the signal, the temporal trajectory of modulation patterns is obtained. Our AMS decomposition uses 17 (Bark-)spectral (50 Hz to 3400 Hz) and 29 modulation-spectral (2 Hz to 30 Hz) bands, resulting in a 493-dimensional representation of the signal.

3. Feature selection and Classification

Feature selection has been pursued with the goal to identify which parts of the amplitude modulation spectrogram are most salient for the detection of speech within realistic acoustic background sounds. Individual frequency/modulation-frequency points are selected from the amplitude modulation spectrogram to maximize classification accuracy. Hence, feature se-
from -20 dB to 20 dB (decibel) and had to be discriminated.

The sequential forward floating search (SFFS, [13]) algorithm is employed. It iteratively adds the most salient feature in each iteration and after each feature inclusion step performs a “backtracking” step in which it is checked whether exclusion of a feature added earlier would lead to an increase in classification accuracy, thereby reducing the risk of local minima in the combinatorial feature space.

The classification back-end employed consists of a standard support-vector-machine classifier [3], which was used with a linear kernel as pilot experiments did not clearly indicate a significant gain with non-linear kernels on these data. Cross-validation folds have been chosen as contiguous parts (and not randomly chosen frames) of the training data in order to avoid artificially high cross-validation accuracy resulting from the presence of low-frequency modulations.

4. Experiments

4.1. Data

The data used for the experiments consists of speech (TIMIT database), background sound (recordings of road traffic noise), and two additional specific noise sources (inside driving car, factory, taken from NOISEX database). About 5 minutes of data were used for training and testing, respectively. Train and test data was disjoint and different speakers were used in the training and testing material. Either clean speech data was used for classifier training (sec. 4.2), or speech was embedded in the background at different long-term signal-to-noise ratios (SNR) from -20 dB to 20 dB (decibel) and had to be discriminated from “pure” background noise (sec. 4.3 and 4.4). The analysis is presented with road traffic background material; similar results have been obtained with other background material (not reported here). The approach of artificially mixing speech and background has been taken since the standard databases at our disposal either do not contain typical background systematically varied at different SNRs, or do not contain unconstrained speech (as in the TI-digits-based Aurora-2 database).

4.2. Discrimination of clean speech vs. background sounds

This section describes simplified experiments in which the classifier is trained to discriminate clean speech (i.e., not embedded in a background) from background sounds alone, a task that is much easier than the full speech-in-background detection task investigated in sections 4.3 and 4.4 below. The performance of the proposed classification scheme is evaluated for different numbers of features. Evaluation is performed on the testing portion of the speech and noise background data sets, and on noise signals from different recordings than trained on in order to test generalization performance.

Fig. 2 displays the results of generalization to unseen data as a function of number of features. Classification accuracy on test data is above 95% already for less than five features, but a higher number of features (between 30 and 40) tends to result in better classification for the factory noise signal that is not similar to the road traffic sound the classifier was trained on.

To illustrate the limitation of a speech detection classifier trained on clean speech data, evaluation is performed for speech mixed with noise signals at different signal-to-noise ratios (SNRs). The resulting classifier accuracy (defined as correctly classified segments relative to total number of segments) is displayed in Fig. 1, right panel. When tested on noisy speech data, the classifier degrades with increasing SNR, reaching chance accuracy at about 0 dB SNR, and classifying the signal as most similar to the background sound class for SNRs below 0 dB SNR, as should be expected. Hence, to achieve reliable detection of speech embedded in noise, the classifier will need to be trained accordingly, as described in the following sections.

4.3. Detection of speech embedded in real background

This section investigates detection of speech embedded in a real acoustic road traffic background. Classifier training is performed for discrimination between speech embedded in the background signal at different SNRs and the pure background signal without the presence of speech.

Results for road traffic noise are displayed in Fig. 3. Cross-validation accuracy during training reaches its maximum typi-
proposed AMS method on average still better than chance. However, even at -20 dB performance is to degrade below 0 dB SNR with a clear deterioration visible performance on test data is very good for positive SNRs. It starts results from training several classifiers on different SNRs. Per-
lows us to compose ROC curves (Fig. 3, right panel, com-
mand in speech. This range coincides in large part with the spectral position of the first for-
peak in the speech modulation spectrum. Our method clearly close to 4 Hz which is generally thought to correspond to the focussed in the modulation frequency range from 2 to 5 Hz, low SNRs (-20 dB, crosses in Fig. 3 center), they are even more focussed in the modulation frequency range from 2 to 5 Hz, close to 4 Hz which is generally thought to correspond to the peak in the speech modulation spectrum. Our method clearly indicates the region of center frequencies from about 550 Hz to about 850 Hz as most salient for the detection task. This range coincides in large part with the spectral position of the first formant in speech.

Training and testing with all combinations of classifiers al-
ways with 10 or less out of 493 features (Fig. 3, left). The first 5 features are already sufficient for very accurate classification. Detection of speech in road noise is highly reliable with cross-validation accuracy exceeding 90% for speech at an adverse level of -10 dB. At high SNRs (20 dB, circles in Fig. 3 center), the 5 most significant features cluster between about 2 and 9 Hz modulation frequency which is compatible with known results [4, 8]. They span a rather broad range of center frequencies. At low SNRs (-20 dB, crosses in Fig. 3 center), they are even more focussed in the modulation frequency range from 2 to 5 Hz, about 850 Hz as most salient for the detection task. This range coincides in large part with the spectral position of the first formant in speech.

4.4. Comparison to ITU G729.B standard

The proposed method has been compared with the standard ITU G729.B voice activity detection (VAD) algorithm on the same test data that has been used in the previous section. Since natively the G729.B works on a frame by frame basis, its speech/non-speech decision has been averaged over 1 sec. intervals to make it comparable to the proposed AMS method. Results are displayed in Fig. 4 for overall accuracy (left, defined as correctly classified segments relative to total number of segments), true positive rate (center) and false positive rate (right) as a function of SNR from -20 dB to 20 dB. Accuracy of the AMS method outperforms G729.B for all SNRs. When decomposing overall accuracy into true positive and false positive rate, it becomes evident that the main drawback of the the G729.B is its high false positive rate: In 60% of the test segments, a pure background signal that does not contain any speech is classified as speech signal. This rate is constant for all SNR because the pure background noise is identical over all test conditions, and the G729.B VAD is not SNR-dependent. In contrast, the AMS method features a significantly lower false positive rate which results in an increased overall accuracy. Also the true positive rate of the proposed method significantly improves on the baseline of the G729.B for all SNRs except the lowest (-20 dB, where true positive rate is at chance level) and highest (20 dB, where the G729.B works very well). These results indicate that the G729.B works very well for the detection of speech in envi-

eronments with a low level of background noise. It fails for the
task of detecting speech embedded in a background predominantly since it too often classifies the background as speech. This fact is most likely induced by the strongly fluctuating characteristics of the real background used. Further, it misses true positives in intermediate SNR regimes (-10 dB, 0 dB, 10 dB), most likely since background fluctuations mask fluctuations of the speech signal.

5. Discussion and conclusion

An approach for the detection of speech embedded in real background sounds has been proposed that is based on the amplitude modulation spectrogram. The accuracy obtained depends on the signal-to-noise-ratio of speech in its background. Detection can be performed with high accuracy for SNRs of near 0 dB, degrades gracefully for lower SNRs and is above chance level for -20 dB.

Feature selection was used to identify the features leading to highest classification accuracy. They have been found to be located in a modulation frequency range from about 2 Hz to about 10 Hz, a range that is known in the literature to be highly relevant for speech processing [4]. For adverse (-20 dB SNR) conditions, the best features are highly focussed near the frequency range of the first formant of speech and in a modulation frequency range between 2 and 5 Hz which is highly compatible with previous literature results from other applications [8].

Compared to the standard energy-based ITU G729.B voice activity detection scheme, the proposed method results in improved performance for all SNR regimes. The G729.B energy-based approach does not perform very well when the target is embedded in real, fluctuating noise backgrounds most likely since (1) presence of the noise signal energy masks energy cues of the speech target and (2) fluctuations in noise signal energy lead to many false detections. This behavior appears to reflect the fact that the G729.B VAD is intended for use in telephony systems with predominantly high SNR speech conditions. Further, the application implies a bias towards speech with ambiguous frames more likely classified as speech rather than noise, resulting in a high false positive rate when tested on entirely non-speech signals. Our method is essentially also looking at energy cues. But through decomposing signal energy into different center-frequency and modulation-frequency bands, and by training a classifier on typical modulation patterns in speech and background, it achieves a significantly higher selectivity for speech.

Our method has been deliberately constructed in such a way that it exploits only modulation information present in the speech and background signal. As a result, signal components that correspond to purely spectral information, e.g., the pitch component, do not contribute towards correct classification. It is expected that by also considering spectral cues, accuracy can be further increased.

The proposed approach of using “meso-scale” modulation features of 1 s length appears promising for identification of speech in real backgrounds. It’s applicability to other types of acoustic objects and backgrounds is currently under investigation [1, 10].

6. Acknowledgements

Research was supported by the EC under the DIRAC integrated project IST-027787 and by DFG International Graduate School for Neurosensory Science and Systems. The authors are grateful for the support of B. Kollmeier and the members of the Med-