Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 690–706
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Robust speech detection in real acoustic backgrounds with perceptually motivated features

Jörg-Hendrik Bach *, Jörn Anemüller, Birger Kollmeier

Medical Physics Department, University of Oldenburg, 26111 Oldenburg, Germany

Available online 21 July 2010

Abstract

The current study presents an analysis of the robustness of a speech detector in real background sounds. One of the most important aspects of automatic speech/nonspeech classification is robustness in the presence of strongly varying external conditions. These include variations of the signal-to-noise ratio as well as fluctuations of the background noise. These variations are systematically evaluated by choosing different mismatched conditions between training and testing of the speech/nonspeech classifiers. The detection performance of the classifier with respect to these mismatched conditions is used as a measure of robustness and generalisation. The generalisation towards un-trained SNR conditions and unknown background noises is evaluated and compared to a matched baseline condition.

The classifier consists of a feature front-end, which computes amplitude modulation spectral features (AMS), and a support vector machine (SVM) back-end. The AMS features are based on Fourier decomposition over time of short-term spectrograms. Mel-frequency cepstral coefficients (MFCC) as well as relative spectral features (RASTA) based on perceptual linear prediction (PLP) serve as baseline.

The results show that RASTA-filtered PLP features perform best in the matched task. In the generalisation tasks however, the AMS features emerge as more robust in most cases, while MFCC features are outperformed by both other feature types.

In a second set of experiments, a hierarchical approach is analysed which employs a background classification step prior to the speech/nonspeech classifier in order to improve the robustness of the detection scores in novel backgrounds. The background sounds used are recorded in typical everyday scenarios. The hierarchy provides a benefit in overall performance if the robust AMS features are employed.

The generalisation capabilities of the hierarchy towards novel backgrounds and SNRs is found to be optimal when a limited number of training backgrounds is used (compared to the inclusion of all available background data). The best backgrounds in terms of generalisation capabilities are found to be backgrounds in which some component of speech (such as unintelligible background babble) is present, which corroborates the hypothesis that the AMS features provide a decomposition of signals which is by itself very suitable for training very general speech/nonspeech detectors. This is also supported by the finding that the SVMs combined with RASTA-PLPs require nonlinear kernels to reach a similar performance as the AMS patterns with linear kernels.

Keywords: Speech detection; Pattern classification; Amplitude modulations; Fluctuating noise; Real-world scenario

1. Introduction

The identification and subsequent segregation of acoustic objects that are embedded in acoustic backgrounds is a task performed exceptionally well by biological auditory systems (Bregman, 1990; Bee and Klump, 2004). While engineering solutions to object detection and classification work comparably well in situations with high signal-to-noise ratios (SNR), satisfactory classification of objects embedded in a background at low SNRs is often at the limits of current technology, in particular with real acoustic backgrounds. Such real-life scenarios are characterized by their non-stationary nature, which results in strong modulations of the background's spectral power density.

The human auditory system is superior to technical systems in that it is able to generalise to unknown acoustical scenes, i.e., it performs well under novel, un-trained scenarios. It is therefore advisable to mimic the relevant

* Corresponding author. Address: Carl-von-Ossietzky-Straße 8-11, 26111 Oldenburg, Germany. Tel.: +49 441 798 3382.
E-mail address: j.bach@uni-oldenburg.de (J.-H. Bach).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.07.003

properties of the human auditory system in order to improve the robustness and performance of artificial systems.

Several studies have indicated that the exploitation of temporal information plays an important role in sound processing both for humans and for animals. Neurophysiological studies in animals (Schreiner and Langner, 1988) as well as magneto-encephalographic studies in humans (Langer et al., 1997) have suggested that amplitude modulations are explicitly coded in the auditory cortex; this motivates incorporating a signal decomposition in terms of spectral *and* modulation frequencies. Kollmeier and Koch (1994), Tchorz and Kollmeier (2003) demonstrated that binaural processing based on a two-dimensional representation of acoustic frequency vs. modulation frequency in each (acoustic) frequency band increases the SNR in those frequency-modulation frequency points relevant for speech. This leads to an improvement in intelligibility (approx. 2 dB) compared to a typical frequency analysis or wideband modulation frequency analysis. Since then, modulation decompositions have been used in various speech-related tasks (Greenberg and Kingsbury, 1997; Mesgarani et al., 2006).

Previous research on acoustic object identification and discrimination encompasses a range of different systems, from simple voice activity detection schemes (VAD) to hearing aid related scene classification. These VADs are based on a feature extraction approach first proposed by Rabiner and Sambur (1975), which extracts the broadband signal energy and the zero crossing rate as a rough measure of the spectral properties. The original approach has since been extended in a variety of ways: one current telephony standard (defined in G729, annex B, see ITU, 1996) uses a VAD that is based on energy, zero crossing rate, and a spectral distortion measure. It performs well within the confines of telephony applications. Even better performance is obtained when the energies associated with different spectral sub-bands are evaluated separately (Marzinzik and Kollmeier, 2002).

More complex methods make use of temporal information as well as spectral measures. For use in hearing aids, for example, explicit modulation decomposition methods have been successfully used for scene classification (Ostendorf et al., 1998). A simpler envelope statistics-based modulation analysis has been put forward by Büchler et al. (2005). In their study, everyday sounds were classified into classes including speech, noise, speech in noise, and music. The evaluation was based on actual recordings, but lacked a systematic robustness analysis; for example the SNR in the recordings varied between +2 dB and −9 dB, but was not systematically controlled. That study tested several classifiers and evaluated which features, from a given set (including envelope modulations), provided optimal results for each classifier. The hit rates for all classes ranged between 78% and 93%.

Mesgarani et al. (2006) used simultaneous spectral and temporal modulation analysis in a feature extraction inspired by biological findings in order to implement an auditory speech/nonspeech detector. Their algorithm consists of strongly auditory motivated steps, including auditory transformed spectrograms (with constant-Q basilar membrane filters and an inner hair cell model) as models for the first auditory processing steps, a lateral inhibition network as auditory nerve model and a wavelet transform with Gabor filters as mother wavelet as a model for the cortical analysis. The Gabor filter acts as a model for spectro-temporal receptive fields, which analyse the input in terms of its spectro-temporal modulations. These features (reduced to less than 33 principle components) increase the robustness against white and pink noise compared to then state-of-the-art approaches by almost 20 dB. The robustness against reverberation is also increased, with the auditory model sustaining approx. 2.5 times the delay times of the other approaches at comparable error rates.

An application of modulation based speech detection as a pre-processing step for automatic speech recognition (ASR) was presented in (Maganti et al., 2007). Their feature front-end extracts amplitude modulations in the range of 2 Hz to 16 Hz in mel-scaled frequency channels. If the smoothed modulation energy in one channel exceeds a fixed threshold, the information in that channel was taken as in indication for the presence of speech. The majority vote over all frequency channels then determined the overall speech/nonspeech decision. Since the application was ASR, the evaluation was performed using meeting room speech recorded at a distance of approx. 1 m to the speaker; i.e., the SNRs were relatively high. The performance of the speech detection was evaluated indirectly by comparing the word error rates of identical ASR systems preceded by different speech segmentation algorithms. The modulation based approach outperformed manual segmentation, zero crossing, and energy-based methods as well as a machine learning-based technique.

Modulation features have also been used in music analysis: Markaki et al. (2008) use an FFT-based amplitude modulation extraction similar to the one presented here to detect singing parts in a particular type of music (Greek Rembetiko). One main effort of their work was to implement higher order singular value decomposition to simultaneously and independently reduce the frequency and modulation frequency space. They found a subspace of 80 (of an original 800) principal components in the frequency-modulation frequency space to be sufficient for optimal performance. Comparing and combining the modulation features (amplitude modulation spectrogram, AMS) with mel-frequency cepstral coefficients (MFCC), their main finding was that AMS and MFCC carry partly complementary information of human (singing) voices.

The present contribution investigates the utility of coupling perceptually motivated acoustic features as a front-end with support vector machines (SVM, Chang and Lin, 2001) as a classification back-end. The experiments presented here aim at systematically evaluating different robustness properties of these features. It has been shown before that drawing inspiration from biological sound processing