Speech / Non-Speech Discrimination for Acoustic Monitoring Considering Privacy Issues

Robert Rehr[1], Stefan Goetze[1], Danilo Hollosi[1], Jens-E. Appell[1] and Jörg Bitzer[1,2]

[1] *Fraunhofer IDMT, Project group Hearing, Speech and Audio Technology (HSA), Oldenburg, Germany*

Email: robert.rehr@idmt.fraunhofer.de

[2] *Jade University of Applied Sciences, Oldenburg, Germany*

Abstract

A reliable decision if acoustic signals contain speech utterances is important in various signal processing problems, such as noise reduction or acoustic monitoring, e.g. in the context of ambient assisted living (AAL) where technical systems are used to unobtrusively support older persons in their daily life. Examples for such systems range from acoustic monitoring of activities of daily living (ADL) to automatic fall detection. Here, ethical issues like privacy protection have to be considered such as deleting contextual information like speech utterance from the data stream.

The aim of this contribution is to find an algorithm, which can safely differentiate between speech and noise events. Since many voice activity detection (VAD) methods often only exploit a single feature within a specialized algorithm, they might be unable to discriminate such events and will erroneously classify all events as speech. Therefore, an approach introduced by Shafiee et al. [1], who used machine learning (ML) based algorithms, is used. We compared the discrimination power of several recently published features. Afterwards the best features are selected and used within a ML framework. The performance of the developed algorithm is compared against various VAD approaches allowing us to present results of first experiments.

Introduction

In this paper different VAD algorithms were analysed which are characterized by low complexity. Therefore they extract simple features on the blocks of the input signal in order to distinguish between speech and non-speech segments. Usually, background noise in the acoustic environment is estimated. This information is often used for a decision rule and a block is tagged as speech if a certain threshold is exceeded. VADs usually have in common that the detection of speech segments is based on a noise estimation and no further information about the speech properties is exploited.

In contrast to that, [1] uses the term speech activity detector (SAD) to distinguish ML based methods from the VADs. These techniques use models to describe different classes allowing the information about speech to be incorporated in the decision process.

In this paper we analyse the performance of selected VAD algorithms and a ML based detection scheme to investigate if ML based techniques yield advantages towards the well-known VAD approaches. These investigations contribute to a more complex approach for acoustic event detection in smart home environments introduced in [2].

Methods

Audio Material

Speech detection algorithms are usually compared using standardized speech databases. However, these databases usually do not contain recordings of realistic household environments. Therefore an own corpus was generated in addition to standard databases. We recorded several scenes in a typical household that always contained a dialogue between a male and a female speaker. In order to recreate the appropriate acoustic environments, typical household activities were performed during the recordings e.g. cooking water, running a microwave or doing the dishes. We used a single microphone which was placed in the far field for a realistic recording of the acoustic environment. Because of using ML based detection methods, the audio material was split into a training and a test set. The spoken texts and the speakers in both sets were different. The audio material was annotated by hand.

Algorithms

In our experiments we found that the VAD algorithms proposed by [3], [4] and [5] are most appropriate for the given problem. We used an improved version of the algorithm [3] where we substituted the proposed noise estimation with an algorithm proposed by [6], [7].

Additionally, we used a simple GMM-based classification scheme as a representative of the ML based algorithms. The system extracts a set of features on each input block. Using GMMs, we calculate a likelihood value for the speech and the noise model which have been previously trained. Afterwards the block is assigned to the class which maximizes this value.

Comparison

In order to guarantee a *fair* comparison, each algorithm was optimized. First, the VADs were adjusted by selecting the best parameters for every test scene. For this purpose a large set of parameter combinations was tested on the audio material. The best parameter set was chosen by means of the detection rate P which is the mean value of the speech hit rate (SHR) and the correct rejection rate (CRR).

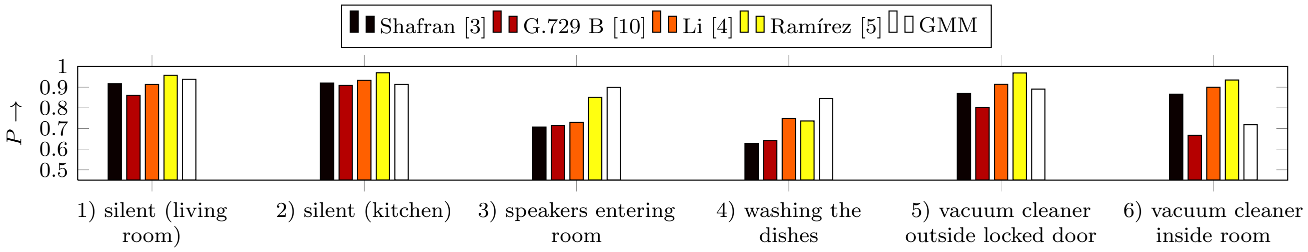

Figure 1: Measured detection rates P of the compared VADs (gray shaded / coloured bars) and the ML based classifiers (none shaded bar) for six scenes of the household corpus

The ML based algorithms were also fitted to the audio material. In previous experiments, which were part of this work, we analysed many features and selected 20 which provided the highest ability to discriminate speech against non-speech. Among the best features are fundamental pitch, spectral flatness, entropy or spectral rolloff, which are mainly described in [8] and [9]. The models consisted of four multivariate gaussians and were trained for speech and background noise. The speech model was created by using utterances in mostly clean conditions, since it reduced the amount of false alarms. The noise model was trained with all noise material available in the training set. Many VADs smooth their speech/non-speech decision by employing so called hangover schemes. No such scheme was used for the ML approach.

Results

We obtain the algorithms' performance by comparing the achieved detection rates, as shown in Figure 1 for selected scenes of the household corpus. In order to facilitate the comparison with findings in other literature, results of the VAD used in [10] were added.

Looking at the first two scenes that do not contain any background noise, we see that all algorithms reach a very high detection rate and thus we can state that speech segments can be easily detected by using any method.

Scene three and four contain many transient non-speech events e.g. door claps, press of a light switch or dish clatter. In such cases the GMM classifier outperforms many VAD algorithms. This is because the ML approach rarely classifies short events as speech because such events do not fit the speech model. In contrary, the VADs are only able to identify deviations within their noise estimation. Therefore, they classify many noise blocks erroneously as speech segments and the detection rate decreases.

The background scenes five and six is rather stationary. Within scene five a vacuum cleaner is placed outside the room behind a closed door. The conversation can be easily understood. Looking at the detection rates, both the VADs and the ML approach are able to detect speech segments correctly. However, the detection rate of the GMM based detection system decreases for lower SNR as the results for scene six indicate. Due to the high noise energy most blocks better fit the noise model even if they contain speech utterances. Hence the ML based system misses many speech segments. On the contrary, the VADs are able to detect the segments correctly because the deviations of the noise estimation, which still can be measured, usually belong to the speech segments.

Conclusion

In many cases, VADs are not able to distinguish between speech and transient noises e.g. a door knock. On the contrary, the GMM based algorithm is able to perform this discrimination at least if the SNR is sufficiently high. Therefore, an ML based algorithm appears to be the more adequate solution. In our application scenario the detection of speech segments is more important during phases of high SNRs where the conversations could be easily understood. However, a rough estimate of the speech segments might be sufficient in low SNRs phases, because the noise itself makes understanding of the conversation more difficult.

References

[1] Shafiee, S., Almasganj, F., Vazirnezhad, B., Jafari, A.: A two-stage speech activity detection system considering fractal aspects of prosody. Pattern Recognition Letters **31**(9) (2009) 936–948

[2] Hollosi, D., Schröder, J., Goetze, S., Appell, J.E.: Voice activity detection driven acoustic event classification for monitoring in smart homes. In: ISABEL Conference, Barcelona, Spain (2010)

[3] Shafran, I., Rose, R.: Robust speech detection and segmentation for real-time ASR applications. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). Volume 1., IEEE (Apr 2003) 432–435

[4] Li, Q., Zheng, J., Tsai, A., Zhou, Q.: Robust endpoint detection and energy normalization for real-time speech and speaker recognition. IEEE Transactions on Speech and Audio Processing **10**(3) (2002) 146–157

[5] Ramírez, J., Segura, J.C., Benítez, C., Torre, Á.d.l., Rubio, A.: Efficient voice activity detection algorithms using long-term speech information. Speech Communication **42**(3-4) (2004) 271–287

[6] Cohen, I., Berdugo, B.: Speech enhancement for non-stationary noise environments. Signal Processing **81**(11) (2001) 2403–2418

[7] Schepker, H.: Weiterentwicklung und Evaluation von Verfahren zur subjektiven und objektiven Bestimmung von Höranstrengung. Bachelorarbeit, Jade Hochschule, Oldenburg, Deutschland (Feb 2011)

[8] Peeters, G.: A large set of audio features for sound description (similarity and classification) in the CUIDADO project (Apr 2004)

[9] Shen, J., Hung, J., Lee, L.: Robust entropy-based endpoint detection for speech recognition in noisy environments. In: International Conference on Spoken Language Processing (ICSLP), Sydney, Australia (Nov 1998)

[10] ITU: G. 729 Annex B, "A silence compression scheme for G. 729 optimized for terminals conforming to recommendation V. 70". International Telecom Union (1996)