Speech Activity Detection for Activity Monitoring using an Embedded Platform

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

[1] *Fraunhofer Institute for Digital Media Technology / Hearing, Speech and Audio Technology, Oldenburg*

[2] *Institute for Hearing Technology and Audiology (IHA), Jade University of Applied Sciences, Oldenburg*

sven.wilksen@idmt.fraunhofer.de

Abstract

The discrimination of speech signals and other signals is needed in different scenarios like hands-free communication systems, hearing aids, speech input systems or acoustic monitoring systems. For the latter one ethical issues like privacy protection usually are very important. In particular, speech passages in audio recordings need to be removed instantaneously in a pre-processing stage such that contextual information cannot be accessed at any time. Therefore, speech needs to be accurately detected in an audio signal in real-time, without compromising the results for the desired event signals Motivated by the successful utilization of voice activity detection (VAD) algorithms in information technology, this paper investigates their suitability and performance in monitoring applications under realistic acoustic conditions. We implemented several suitable VAD algorithms on an embedded platform. This was done to evaluate the algorithms according to their CPU and memory consumption and to allow the development of a compact, ambient and economically efficient electronic device.

Introduction

Acoustical monitoring systems can be used to support older people in their living environment e.g. by detecting shouts for help and other possibly dangerous situations. Such monitoring systems need to be compact in size and economically efficient, i.e. have long battery life-cycles, as they are destined to be used in private homes. Privacy protection is one of the most important issues in the context of home monitoring. The transmission or recording of speech is often not wanted and thus has to be prevented. For this purpose, algorithms for robust voice activity detection are needed. These algorithms must be able to separate speech and speech pauses also in noisy household environments.

The efficiency and compactness requirements for monitoring system are fulfilled by embedded computers, although these systems tend to have lower processing power than conventional personal computers. VAD algorithms are commonly used in telecommunication applications to separate speech and speech pauses on systems with low processing power, e.g. mobile phones. The classification results of VAD algorithms can also be used as control signals to stop signal transmission or to trigger speech recognition stages.

These kinds of algorithms are usually designed to be used in communication applications with speech recorded in near-field. The algorithms are generally not evaluated with speech in household-noise environments. In this paper, we investigate the detection performance of VAD algorithms for speech in household-noise environments with audio signals recorded in far-field.

We chose the BEAGLEBOARD [7] single-board computer as embedded platform, since LINUX distributions are available for this system which allow easy and platform-independent software development. The BEAGLEBOARD features a 720 MHz ARM Cortex A9 CPU, 256 MB memory and an onboard audio interface. With all necessary hardware on board and an energy consumption of less than 2 Watts, this platform matched all of our requirements.

A selection of three VAD algorithms was implemented both in real-time on the BEAGLEBOARD and in a file I/O framework on a conventional personal computer for evaluation purposes: Ramirez et al. 2004 [2], Marzinzik and Kollmeier [5] and Shafran and Rose [3] with the noise estimation from Cohen and Berdugo [4]. Each algorithm was implemented in C++ following the algorithm description in the corresponding articles. Additionally, the VAD module output of the ITU G.729 Annex B speech coder [6] was chosen as our test-reference for the detection performance measurements. To ensure comparability to other studies, the reference implementation of the ITU was used for the evaluation.

Methods

The suitability of a VAD algorithm for usage in home monitoring applications is evaluated in terms of detection performance as well as CPU load and memory consumption on the embedded platform.

The detection performance of speech in household-noise environments is measured by observing the speech hit rate (SHR), determining the percentage of correctly classified speech segments and the false alarm rate (FAR), which determines the percentage of noise segments classified as speech. A third measure P, the so called detection performance

$$P = \beta_P \cdot \mathrm{SHR} + (1 - \beta_P) \cdot (1 - \mathrm{FAR}) \qquad (1)$$

is computed from the SHR and FAR. The weighting parameter was chosen to be $\beta_P = 0.6$ for our evaluations to focus on higher speech hit rates during optimization and evaluation of the algorithms. These measures are computed by comparing the output of a VAD algorithm against the ground truth annotation of the used test signal.

Audio material

To evaluate the detection performance, audio material matching the target application was needed. Therefore we recorded thirteen scenes with dialogs between two or three people in typical household situations, like speech in absence of noise, operating noises of vacuum cleaners, microwaves etc.

Optimization

Each algorithm was optimized to achieve high detection rates in every test signal. The optimization was done using a brute force method, calculating the performance measure P for a large amount of algorithm configurations and picking the configuration with the highest performance P for each scene. Additionally, the algorithms were optimized over all scenes using one single configuration to simulate realistic operation conditions, where the algorithms are used with fixed parameters.

Results

Speech detection performance

The mean detection performance measures over all scenes are shown in Table 1. It is clearly visible, that the highest detection performance is achieved by the VAD algorithm of Ramirez et al. This algorithm has also the lowest mean FAR. The second best algorithm has the highest FAR. The reference algorithm (G.729 B) shows the poorest detection performance of speech in the analyzed household-noise situations. It should be noted, that this algorithm was not individually optimised for each situation to deliver a test reference. When using a single configuration over all scenes for each algorithm, the detection performances are reduced and only the algorithm of Ramirez et al. is notably better than the reference algorithm. The other algorithms show poorer detection performances than the G.729 B VAD.

Table 1: Mean detection performance over all scenes.

Algorithm	P	SHR	FAR
Ramirez et al.	0.89	0.95	0.2
Marzinzik & Kollmeier	0.76	0.82	0.32
Cohen & Shafran	0.8	0.94	0.41
G.729 B	0.74	0.77	0.29

Further investigations showed that all algorithms tend to have rather poor detection performances for speech in presence of transient and tonal noises. Instantaneous increases of the background noise, such as switching on a vacuum cleaner have a great impact on the detection performance of the VAD algorithm of Ramirez et al. Sudden rises of the background noise are incorrectly classified as speech and the classification does not change until the noise level lowers. Because the noise estimation is only active during speech pauses, miss-classifications during level changes of the background noise lead to continuous miss classifications as the estimated noise is part of the VAD decision and the decision threshold is adapted to the estimated noise energy.

Resource usage

The CPU and memory usage of each algorithm is shown in Table 2. The system load generated by the algorithms is always lower than 17 %. The lowest system load is generated by the VAD algorithm of Ramirez et al. The memory consumption is dominated by the real-time framework itself, as further investigations showed. The differences in memory usage between the algorithms are lower than 0.1 % of the overall available memory. As the standard LINUX system tool `top` was used, the *resolution* of the memory consumption is limited. Further investigations of the real-time framework running without algorithm showed that the CPU load is always below 5% with sampling frequencies up to 22050 Hz.

Table 2: CPU and memory consumption of the algorithms in real time operation.

Algorithm	CPU usage	Memory consumption
Ramirez et al.	5.25 %	0.9 %
Marzinzik & Kollmeier	7.56 %	0.9 %
Cohen & Shafran	16.52 %	0.9 %

Conclusion

The VAD algorithm of Ramirez et al. showed the best performance in our tests on speech activity detection in household-noise environments, both in terms of detection performance and low resource usage on the embedded platform. However, although this algorithm has a quite good detection performance, it can be concluded that speech activity detection with VAD algorithms not always leads to sufficiently reliable results for all investigated household-noise situations. If the algorithm of Ramirez et al. can be further improved for transient noises or fast raising noise levels to be suitable as a preprocessing stage in home monitoring systems.

References

[1] Hollosi, Danilo; Schröder, Jens; Goetze, Stefan; Appell, Jens-E.: Voice Activity Detection Driven Acoustic Event Classification for Monitoring in Smart Homes, 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies, Rome, Italy, November 2010

[2] Ramirez, J. et al.: Efficient voice activity detection algorithms using long-term speech information, Elsevier, Speech communication 42 (2004), 271-287

[3] Shafran, I. and Rose, R.: Robust speech detection and segmentation for real-time ASR applications, Proceedings of ICASP 2003, IEEE

[4] Cohen, I. and Berdugo, B.: Noise estimation by minima controlled recursive averaging for robust speech enhancement, IEEE Signal Processing Letters 9 (2002), 12-15

[5] Marzinzik, M. and Kollmeier, B.: Speech pause detection for noise spectrum estimation by tracking power envelope dynamics, IEEE Transactions on Speech and Audio Processing No. 10 (2002), 109-118

[6] ITU, A silence compression scheme for G. 729 optimized for terminals conforming to Recommendation V. 70, 1996

[7] BEAGLEBOARD project homepage, URL: `http://www.beagleboard.org`