SPEECH SEGMENTATION IN NOISY STREET ENVIRONMENT

Jaroslaw Baszun
Department of Real Time Systems, Faculty of Computer Science
Bialystok University of Technology, Wiejska Str. 45A, 15-351 Bialystok, Poland

Keywords: Voice activity detector.

Abstract: Two voice activity detectors for speaker verification systems were compared in this paper. The first one is single-microphone system based on properties of human speech modulation spectrum i.e. rate of power distribution in modulation frequency domain. Based on the fact that power of modulation components of speech is concentrated in a range from 1 to 16 Hz and depends on rate of syllables uttering by a person. Second one is two-microphone system with algorithm based on coherence computation. Experiments shown superiority of two-microphone system in case of voiced sounds in background.

1 INTRODUCTION

Reliability is the most important issue in practical applications of speech-based applications including speech recognition, speaker verification or speech encoding. The main function of a voice activity detector (VAD) is to indicate speech presence and provide delimiters for the beginning and end of speech segment to facilitate speech processing. Traditional VAD's have relied on the observation that noise is usually stationary or slowly-varying and can be estimated during speech pauses. In devices working in real acoustic environments like streets in centers of big cities this assumption is not true. In such environments many types of noises interfere with speech and reduce recognition performance.

Most single-microphone systems are based on identifying pauses between speech and computing noise estimate in the pauses. The problem is that the noise estimate is not updated during the speech. Such solution (Sovka and Pollak, 1995) work well in case of stationary and slowly-varying noise, but gives false response to fast time-varying noise. Another problem is that noise estimate depends on performance of the voice activity detector. If the noise is rapid time-varying or speech-like sound the noise estimate is not properly updated and the system fails.

Some techniques of computing noise estimate like minimum statistics (Martin, 2001) or use of nonlinear estimate of noise power (Doblinger, 1995) allows overcome problem of noise estimation, but fails in case of speech-like noise. An alternative approach shown in this paper is based on filtration of spectral envelopes of speech signal split into number of bands. This method allows for effective continuous noise estimation and has good performance in non-stationary noise. Based on the fact that power of modulation components of speech is concentrated in a range from 1 to 16 Hz (Houtgast and Steeneken, 1985) and depends on rate of syllables uttering by a person, it is possible to separate speech like sounds from noises (Hermansky and Morgan, 1994)(Atlas and Shamma, 2003). This separation can be carried out by bandpass filtration of spectral envelops.

Single-microphone system was compared to two-microphone system with algorithm based on coherence computation. Two-microphone solution exploits the physical characteristics of real noises which are globally diffused. This implies a weak spatial coherence of such sources in compare to almost punctual source speech signal such like a person talking from short distance in front of microphones.

A number of experiments were carried out to compare both systems and asses their usefulness for intelligent cash machine working in noisy street environment.

Baszun J. (2007).
SPEECH SEGMENTATION IN NOISY STREET ENVIRONMENT.
In *Proceedings of the Second International Conference on Signal Processing and Multimedia Applications*, pages 422-427
DOI: 10.5220/0002139704220427

2 SINGLE-MICROPHONE SYSTEM BASED ON FILTERING OF SPECTRAL ENVELOPES

The first considered system exploits modulation spectrum properties of human speech. It is know that low-frequency modulations of sound are the carrier of information in speech (Drullman et al., 1994)(Elhilali et al., 2003). In the past many studies were made on the effect of noise and reverberation on the human modulation spectrum (Houtgast and Steeneken, 1985)(Houtgast and Steeneken, 1973) usually described through modulation index (MI) as a measure of the energy distribution in modulation frequency domain i.e. normalized power over modulation for a given frequency band at dominant modulation frequencies of speech. MI vary between analysis frequency bands. The corrupting background noise encountered in real environments can be stationary or changing usually different in compare to the rate of change of speech. Relevant modulation frequency components of speech are mainly concentrated between 1 and 16 Hz with higher energies around $3-5$ Hz what corresponding to the number of syllables pronounced per second (Houtgast and Steeneken, 1985), see Fig. 1. Slowly-varying or fast-varying noises will have components outside the speech range. Further, steady tones will only have MI constant component. System capable of tracking speech components in modulation domain are very promising in many fields of speech processing (Thompson and Atlas, 2003)(Hermansky and Morgan, 1994)(Baszun and Petrovsky, 2000)(Mesgarani et al., 2004).

2.1 System Description

The idea of the system was based on work on speech enhancement systems (Baszun and Petrovsky, 2000) based on modulation of speech and its version utilizing Short Time Fourier Transform (STFT) analysis was detailed described in (Baszun, 2007).

The block diagram of modified system was shown in Fig. 2. In this approach signal from microphone with sampling frequency 16 kHz is split into $M = 64$ frequency bands using polyphase uniform DFT analysis filter. This allows for better separation of adjacent channels in compare to analysis based on STFT. Next amplitude envelope is calculated for first 33 bands except first band corresponding to frequencies below 250 Hz in speech signal.

Amplitude envelope is summed for all bands and

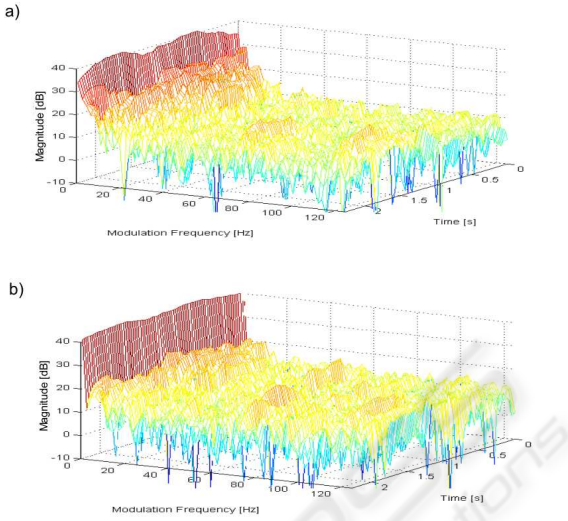

Figure 1: Changes of modulation spectrum in passband from 2375 Hz to 2625 Hz: a) clear speech; b) speech with noise SNR = 10 dB.

filtered by passband IIR filter with center frequency 3.5 Hz and frequency response shown in Fig. 3. The output of the filter is half-wave rectified to remove negative values from output of the filter. The following computation is carried out on the filtered and not filtered envelope:

$$S(nM) = \frac{Y'}{Y - mean(Y) - Y' - mean(Y')} \quad (1)$$

Above parameter is an estimate of speech to noise ratio of analyzed signal. Mean value of filtered and nonfiltered envelope is computed based on exponential averaging with time constant approximately 1 s. The square of this estimate is used as a classification parameter for voice activity detector. Speech decision is based on comparison between classification parameter and the threshold computed based on the following statistics (Sovka and Pollak, 1995):

$$Thr = mean(d) + \alpha \cdot std(d) \quad (2)$$

where d is a classification parameter and α controls confidence limits and is usually in the range 1 to 2, here was set to be equal 2. Both mean value and standard deviation is estimated by exponential averaging in pauses with time constant $a = 0.05$. Frame is considered to be active if value of classifier is greater than threshold. To avoid isolated errors on output caused by short silence periods in speech or short interferences correction mechanism described in (El-Maleh and Kabal, 1997) was implementing. If current state generating by the VAD algorithm does not differ

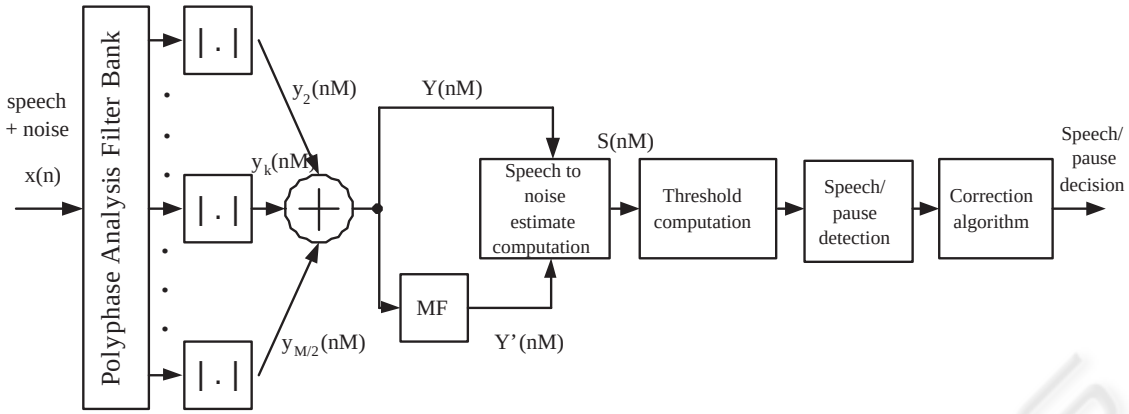

Figure 2: Block diagram of voice activity detector based on filtering of spectral envelopes.

from *n* previous states then current decision is passed to detector output otherwise the state is treated as a accidental error and output stays unchanged.

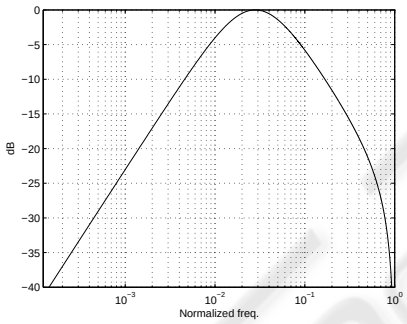

Figure 3: Magnitude frequency response of modulation filter (MF).

3 TWO-MICROPHONE SYSTEM

Two-microphone VAD system exploits the physical characteristics of real noises which are globally diffused. This implies a weak spatial coherence of such sources in compare to almost punctual source speech signal such like a person talking from short distance in front of microphones (Martin and Vary, 1994). The system was motivated by paper (Guerin, 2000). Two omnidirectional matched microphones were used. Microphones were mounted in distance of 40 cm. The speaker was located in 20 to 40 cm from the center of two microphones. In Fig. 4 block scheme of the system was shown. Sampling frequency for two channels was 16 kHz, 16 bit/sample. Magnitude squared coherence function was computed for 512 points without overlapping. This function between two wide-sense stationary random processes x and y is defined as follows (Carter, 1987):

$$\gamma_{xy}(\omega) = \frac{P_{xy}(\omega)}{\sqrt{P_{xx}(\omega)P_{yy}(\omega)}} \tag{3}$$

In practice estimate of magnitude-squared coherence (MSC) is used

$$C_{xy}(k) = \frac{|P_{xy}(k)|^2}{P_{xx}(k)P_{yy}(k)}, \quad 0 \le C_{xy}(k) \le 1 \tag{4}$$

where P_{xx} and P_{yy} are power spectral density of x and y and P_{xy} is the cross power spectral density of x and y. It is assumed that signals from two microphones consists of speech signal and additive noise:

$$\begin{aligned} x(k) &= s_x(k) + n_x(k), \\ y(k) &= s_y(k) + n_y(k), \end{aligned} \tag{5}$$

were s_x and s_y are speech and n_x and n_y are noise components. It was (Martin and Vary, 1994) proved that in most situations noise components are not correlated to each others and speech components are strongly correlated. So in speech segments magnitude squared coherence function is almost equal 1 and in other case its value is near zero. To estimate the magnitude squared coherence function Welch's averaged periodogram method was applied in the system. Next magnitude squared coherence function is integrated, for discrete signal the following approximation was applied:

$$M_\gamma = \frac{1}{N} \sum_{k=k_{min}}^{k_{max}} C_{xy}(k) \tag{6}$$

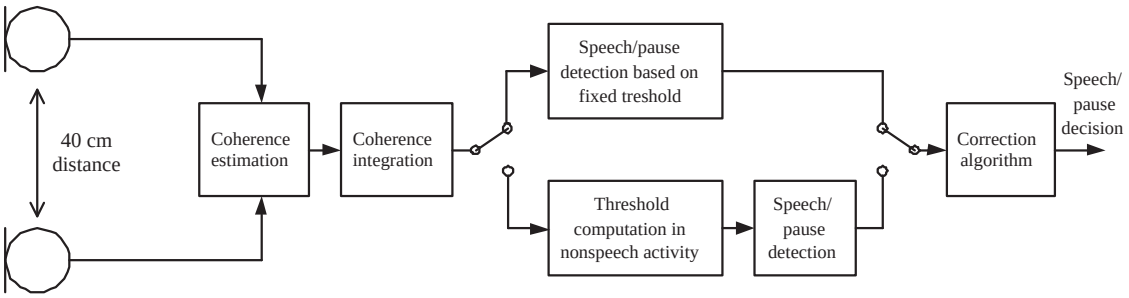

Figure 4: Block diagram of voice activity detector based on coherence estimation.

where $k_{min} = 9$ which means that approximation is not computed for frequencies below 250 Hz where noise is usually strongly correlated for assumed distance of 40 cm between two input microphones (Martin and Vary, 1994). In this algorithm short interferences correction mechanism identical to single-microphone system was implementing.

4 EXPERIMENTS

Algorithms were implemented in Matlab environment. All experiments were carried out with real signals recorded in the noisy street, suburbs of the city (bird chirps and dogs barking) and in classroom. Stereo recordings were made using a pair of matched omnidirectional microphones placed in distance of 40 cm. Speech sequences in collected database were manually marked.

4.1 Results

A few different acoustical situations were investigated. In Fig. 5 recording was made in the street, speech is corrupted by nonstationary noise. A better performance was obtain for single microphone system. In Fig. 6 recording is made three meters from passing cars. The results for single-microphone system is generally better but false detections occurs.

In Fig. 7 recording was made far from noisy street but dog barking is audible between speech segments and during speech segments. In such situation single-microphone system is not able to distinguish speech from dog barking. Two-microphone system deals with this problem because barking picked up by microphones is not correlated.

In Fig. 8 the same situation like in Fig. 7 is shown for birds chirp. Single-microphone system is not able by simple filtration of spectral envelopes to separate such sounds from speech.

In Fig. 9 recording was made in classroom 40 m^2 floor space in situation of cocktail party effect. Single-microphone system give number of false speech detection decisions - separation in modulation domain fails in this case. Coherence based detector is not able to trace all speech segments because SNR of the signal is less than 0 in this recording.

Experiments shown that single-microphone system was not able to distinguish speech from such sounds like dog barking or bird chirps in background and cocktail-party effect. On the other hand single-microphone system has generally better performance in presence of broadband noise.

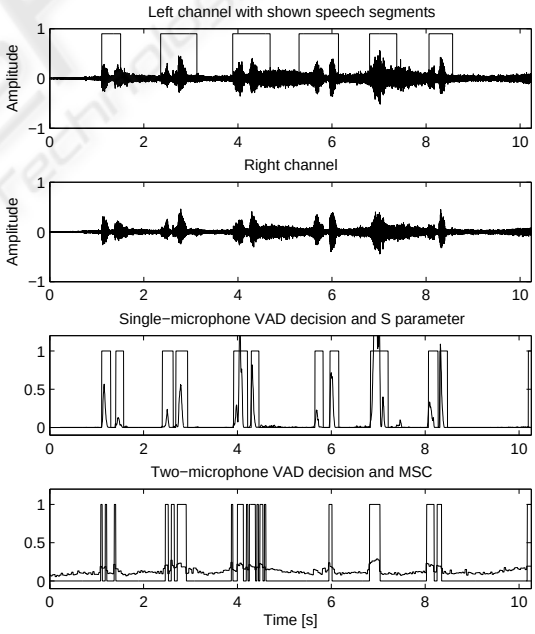

Figure 5: Effects of speech detection in case of speech with nonstationary street noise.

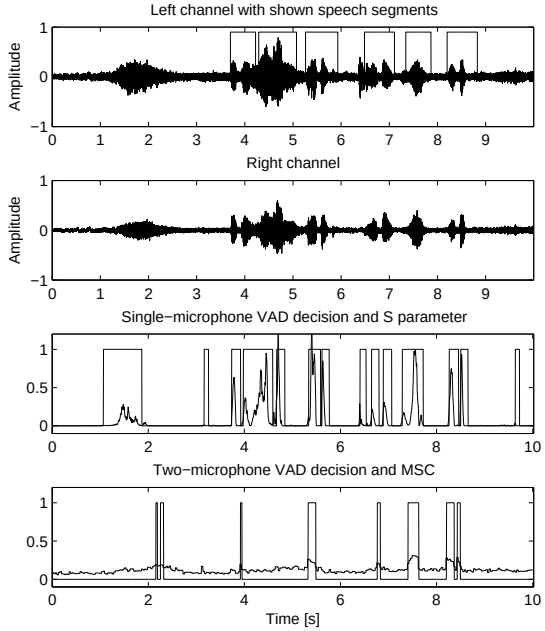

Figure 6: Effects of speech detection in case of very strong noise from passing cars.

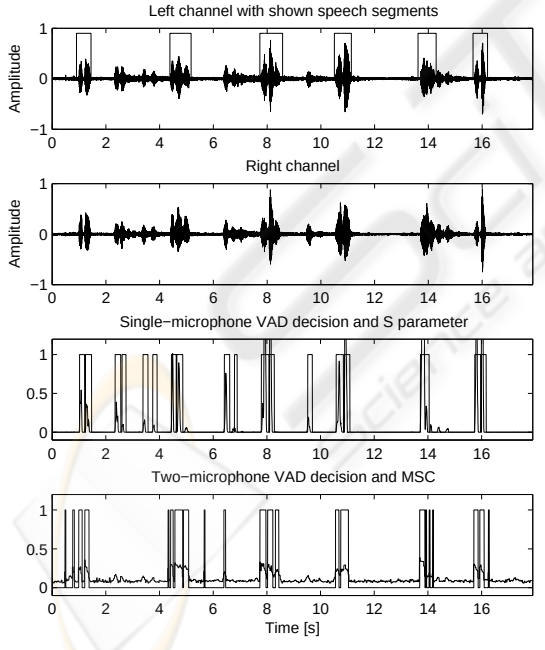

Figure 7: Effects of speech detection in case of speech mixed with dog barking in background.

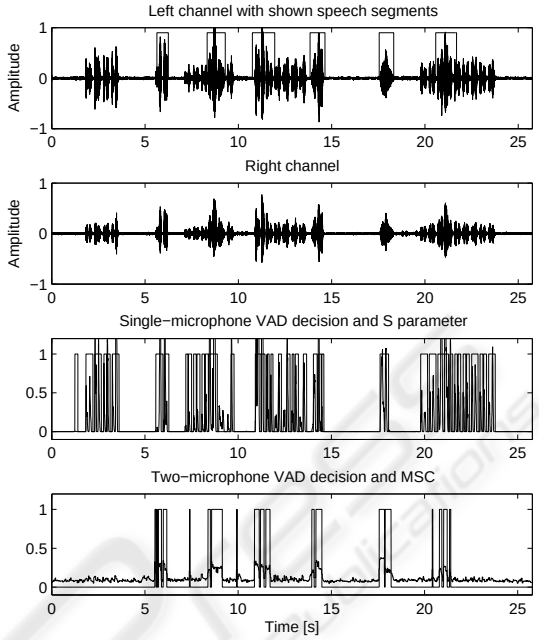

Figure 8: Effects of speech detection in case of speech mixed with bird chirps in background.

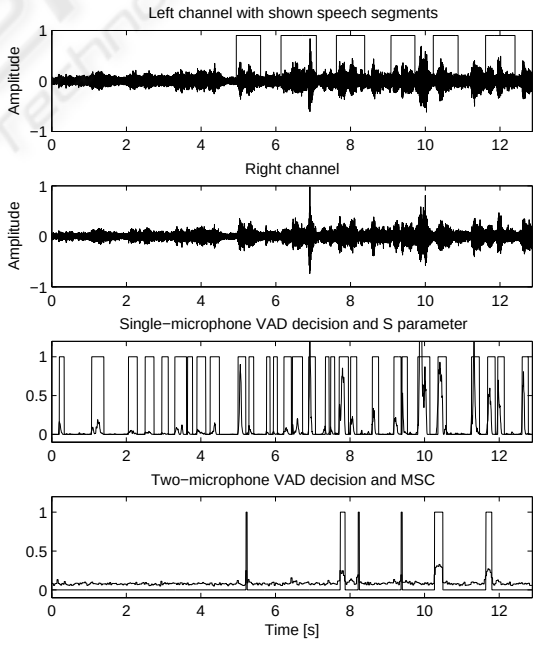

Figure 9: Effects of speech detection in case of cocktail party effect in classroom.

5 CONCLUSION

Two voice activity detectors for speaker verification system were described and compared in this paper. The first one is single-microphone system based on properties of human speech modulation spectrum.

The second one is two-microphone system based on coherence function. Experiments shown better performance of single-microphone system in case of unvoiced background sounds like passing cars. In case of voiced sounds in background simple filtration of modulation components was insufficient to discriminate speech from voiced sounds and coherence based system gives much less false speech detection decisions. Future work will be concentrated on application of coherence function into modulation domain.

ACKNOWLEDGEMENTS

The work presented was developed within VISNET 2, a European Network of Excellence (http://www.visnet-noe.org), funded under the European Commission IST FP6 Programme.

REFERENCES

Atlas, L. and Shamma, S. (2003). The modulation transfer function in room acoustics as a predictor of speech intelligibility. *EURASIP Journal on Applied Signal Processing*, 7:668–675.

Baszun, J. (2007). Voice activity detection for speaker verification systems. In *Joint Rougth Set Symposium*, Toronto, Canada.

Baszun, J. and Petrovsky, A. (2000). Flexible cochlear system based on digital model of cochlea: Structure, algorithms and testing. In *Proceedings of the 10th European Signal Processing Conference (EUSIPCO 2000)*, pages 1863–1866, Tampere, Finland. vol. III.

Carter, G. C. (1987). Coherence and time delay estimation. *Proceedings of the IEEE*, 75(2):236–254.

Doblinger, G. (1995). Computationally efficient speech enhancement by spectral minima tracking in subbands. In *Proceedings of the 4th European Conference on Speech Communication and Technology*, pages 1613–1516, Madrit, Spain.

Drullman, R., Festen, J., and Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.*, (2):1053–1064.

El-Maleh, K. and Kabal, P. (1997). Comparison of voice activity detection algorithms for wireless personal communications systems. In *Proceedings IEEE Canadian Conference Electrical and Computer Engineering*, pages 470–473.

Elhilali, M., Chi, T., and Shamma, S. (2003). A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. *Speech Communication*, 41:331–348.

Guerin, A. (2000). A two-sensor voice activity detection and speech enhancement based on coherence with additional enhancement of low frequencies using pitch information. In *EUSIPCO 2000*, pages 178–182, Tampere, Finland.

Hermansky, H. and Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):587–589.

Houtgast, T. and Steeneken, H. J. M. (1973). The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acustica*, 28:66.

Houtgast, T. and Steeneken, H. J. M. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.*, 77(3):1069–1077.

Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, 9:504–512.

Martin, R. and Vary, P. (1994). Combined acoustic echo cancellation, dereverberation and noise reduction: a two microphone approach. *Ann. Telecommun.*, 49(7-8):429–438.

Mesgarani, N., Shamma, S., and Slaney, M. (2004). Speech discrimination based on multiscale spectro-temporal modulations. In *ICASSP*, pages 601–604.

Sovka, P. and Pollak, P. (1995). The study of speech/pause detectors for speech enhancement methods. In *Proceedings of the 4th European Conference on Speech Communication and Technology*, pages 1575–1578, Madrid, Spain.

Thompson, J. and Atlas, L. (2003). A non-uniform modulation transform for audio coding with increased time resolution. In *ICASSP*, pages 397–400.