# Dual-Microphone Speech Extraction from Signals with Audio Background

Mariusz Ziółko, Bartosz Ziółko, Rafał Samborski
Department of Electronics
AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Kraków, Poland
{ziolko,bziolko}@agh.edu.pl, sambo@tlen.pl

## Abstract

*This paper presents a model of a system which is able to support Police investigations. The listening in system acquires audio signals from two microphones located around one meter to each other. System is designed to record suspect conversations under disadvantageous conditions like the loud music or the car machine noise. The presented method uses the differences between records obtained from two microphones. Different delays and different spectra separates the conversation signal from the background. The crosscorelation analysis and the band-pass filters were applied to realise the suggested algorithm.*

## 1. Introduction

Listening in systems are one of the most efficient and cheapest sources to provide proofs of crimes for Police and homeland security forces investigations [4]. The criminals and terrorists use different methods to hide their conversations. Some of them are very simple in their nature, like turning on a radio close to conversing people. Such methods are effective enough only in the case when an officer is not supported by a speech enhancement system designed for this aim.

Dual-microphone and generally, microphone arrays applied in need of speech enhancement is a well defined field with several methods: beamforming [3], superdirective beamforming [5], postfiltering [1] and phase based filtering [2, 6]. However, in all solutions known to authors it focuses on solving a problem of a random background noise caused by environment where recording takes place. Our case is different in this aspect, because the noise was added intentionally by a conversing human to degradate quality of recordings as much as possible. What is more, the microphones have to be hidden from speakers and in positions where it was simply possible to put a tapping device. This is much different to scenarios typical for information centres or conference rooms. Several efficient methods including phase-based filtering, which is a form of time-frequency masking (PBTFM) [6] require speaker's position to be known. It is all not possible in our scenario because the speakers are in their houses, cars or jail cells and the microphones are listening in devices.

A simple, non-reverberant model for the signals observed by two microphones is given as

$$\begin{aligned} s_1(t) &= s(t) + n_1(t), \\ s_2(t) &= s(t-\tau) + n_2(t), \end{aligned} \tag{1}$$

where $s(t)$ is the speech signal, and delay $\tau$ is caused by a distance longer for the second microphone. Signals $n_1(t)$ and $n_2(t)$ represent microphone noise and environmental noise especially added by conversing persons. It can be a noise of a car machine or a radio, including other human voices like broadcasted news.

The paper is divided as follows. Section 2 gives details on the recording scenario we consider. Section 3 provides mathematical formulae describing calculations we conduct in order of hidden speech extraction. Section 4 covers details on tests and results we achieved. The paper is summed up with conclusions.

## 2. Problem Description

The problem of separating a conversation from the audio signal is depicted in Fig. 1. The audio signals are acquired by two hidden microphones. There are two speaking persons who use a distracting signal, like music from a radio, to block off understanding the content of their conversation. In order to proceed with detecting speech signal from the noised signals recorded by two microphones, at least distances $a \neq b$ or $c \neq d$ must be kept. The difference between these distances can be relatively small. To prove it, let us assume the sampling frequency 44 100 Hz. Then a difference in time between two samples relates to a distance

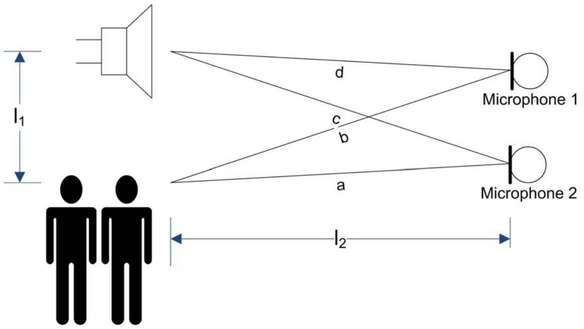
$$\rho = v\Delta t \qquad , \tag{2}$$

978-0-7695-3762-7/09 $26.00 © 2009 IEEE
DOI 10.1109/IIH-MSP.2009.34

IEEE computer society

where $v$ is the sound velocity and $\Delta t$ is the sampling density. For values $v = 330\frac{m}{s}$ and $\Delta t = 23\mu s$ one obtains $\rho \approx 7.5mm$. For a real application, we need at least a difference of around ten samples between signals from both microphones to proceed. This gives a few centimetres as a necessary difference in the distance.

**Figure 1. Dual-microphone scenario of listening in to a conversation where source of a distracting signal, like a radio was used to hide content of the conversation.**



The algorithm described below uses the differences in lengths between distances and additionally differences in frequency bands: higher for music signal and lower for speech, both detectable for an ear. The crosscorrelation and bandpass filters were applied to utilise the properties mentioned above.

## 3. Speech Extraction Algorithm

Let us assume two input signals: $s_{m1}^{in}$ detected in the first microphone and $s_{m2}^{in}$ in the second one. Both of them are filtered with a band-pass filter (BPF) to obtain signals $s_{m1}^{out}$ and $s_{m1}^{out}$. We conducted our experiments on three BPFs: from 7 to 10 kHz, from 4 to 7 kHz and from 4 to 10 kHz and a high-pass filter (HPF) above 4 kHz.

The crosscorrelation for signals obtained from both microphones must be computed to find the difference $\tau_c v$ between paths from the distraction signal source to both microphones. If such a difference exists, then the crosscorrelation $c(\tau_c)$ obtains its maximal value for $\tau_c$, i.e.

$$c(\tau_c) = \max_{\tau} \left[ \sum_n s_{m1}^{out}(n-\tau) s_{m2}^{out}(n) \right] \quad . \tag{3}$$

This is an important information which enables one to remove the audio background signal. Then the speech signal can be found as

$$s_{speech}(n) = s_{m1}^{in}(n-\tau_c) - k\, s_{m2}^{in}(n) \tag{4}$$

where

$$k = \sqrt{\frac{\sum_n (s_{m1}^{out}(n))^2}{\sum_n (s_{m2}^{out}(n))^2}} \tag{5}$$

is an amplification. To enhance voices of both speakers, values $\tau_i$ should be found for which autocorrelation for signal (4) obtains the local extreme values (each for another speaker), i.e.

$$\{\tau_i\}_{i=1}^{2} = \arg\max_{\tau} \left[ \sum_n s_{speech}(n-\tau) s_{speech}(n) \right] . \tag{6}$$

Finally, the enhanced speech of $i$-th speaker can be obtained from the relation

$$s_i = s_{speech}(n-\tau_i) - s_{speech}(n). \tag{7}$$

## 4. Testing and Results

The efficiencies of three BPFs (band-pass from 4 to 10 kHz, from 4 to 7 kHz and from 7 to 10 kHz) and the HPF above 4 kHz were tested. We did not find any important difference in the system behaviour for different BPFs. The reason is that the small part of speech signal energy is located in frequency band above 4 kHz.

**Figure 2. Values of crosscorrelation (3) as a function of a time shift.**

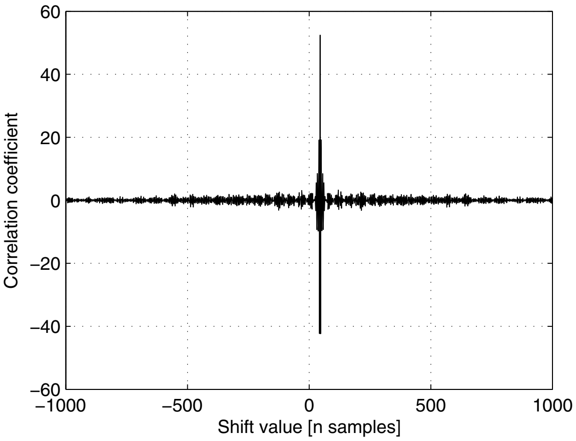
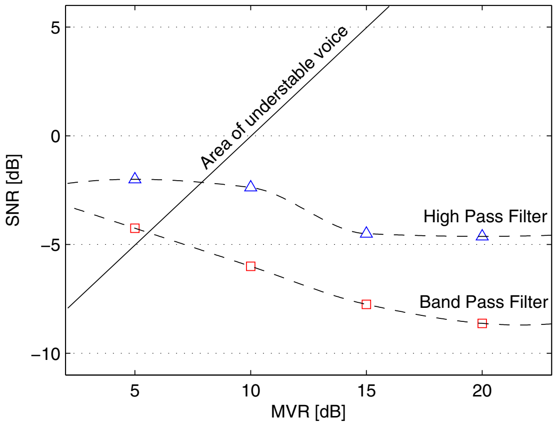


Fig. 3 presents conditions which have to be met to find the correlation and separate a conversation from the recorded signal. The results of BPFs and a HPF were compared. The exact value of the bands in BPFs (4-10 kHz, 4-7 kHz and 7-10 kHz were tested) did not have impact on the results. However, applying HPF gave worse results than BPFs.

In the presented example, the correct shift can be found for a conversation signal hidden 10 dB under music signal

**Figure 3. Comparison of applying two different filters (a BPF from 4 to 7 kHz and a HPF above 4 kHz) based on Matlab simulation. It presents SNR and music to voice ratio (MVR) for which a proper correlation can be found what leads to speech separation from recorded signals. A threshold of conditions for which the speech signal can be understood can be given as 10 dB under a sum of noise and the distracting signal.**



with SNR = -6 when a BPF was applied. This results are for a simplified simulation. The simulation did not cover acoustic effects due to some barriers in the space, especially walls, which, in example, can cause reverberation or several other acoustic issues which can modify a signal.

A higher power of music signal in comparison to speech signal makes it easier to find the correlation which allows to separate conversation from the recorded signal. It is counter-intuitive, but the method starts with applying (3), where it does not matter how strong the speech signal is. This is why such anomaly appear for the range of parameters we are interested in.

White noise was added ten times for each configuration of separate tests, apart from speech and music signal to evaluate the method further. The presented results are mean values for ten experiments with different random noise for a BPF from 4 to 7 kHz. It is possible to find the correlation for a negative SNR, because noise is not correlated in both microphones.

The shift $\tau_c$ between signals was 45 samples with sampling frequency 44 100 Hz. In result, according to (2), music was detected in one microphone around 1 ms faster than in the other one.

## 5. Conclusions

The scenario assumptions are that there are very few possible localisations of microphones and they have to be hidden as listening in devices. The disruptive signal (in example, music from a radio) is added intentionally by conversing speakers to hide the speech content, along with noise. The presented method of signal analysis from two microphones was found successful in recovering conversation from a signal with very low signal to noise ratio and music to voice ratio. BPFs with a band which is hearable for a human, but above speech frequencies were found very successful. Further investigations on more records with numerical evaluating will be conducted.

## 6. Acknowledgements

This work was supported by MNISW OR00001905.

## References

[1] Y. M. C. Marro and K. U. Simmer. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Trans. Speech, Audio, Signal Process.*, 6:240259, 1998.

[2] P. A. A. S. D. Halupka, A. S. Rabi. Low-power dual-microphone speech enhancement using field programmable gate arrays. *IEEE Transactions on Signal Processing*, 55(7):3526–3535, 2007.

[3] G. DeMuth. Frequency domain beamforming techniques. *Proc. IEEE Int. Conference on Acoustics, Speech, Signal Processing*, 2:713–715, 1977.

[4] B. Hołyst. *Kryminalistyka (Eng. Criminology)*. PWN, 1973.

[5] K. D. K. J. Bitzer, K. U. Simmer. Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement. *Proc. IEEE Int. Conference on Acoustics, Speech, Signal Processing*, 5:29652968, 1999.

[6] G. Shi and P. Aarabi. Robust digit recognition using phase-dependent time-frequency masking. *Proc. IEEE Int. Conference on Acoustics, Speech, Signal Processing (ICASSP), Hong Kong*, page 684687, 2003.