

## SPEECH EXTRACTION FROM JAMMED SIGNALS IN DUAL-MICROPHONE SYSTEMS

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

### ABSTRACT

This paper presents two different methods of speech extraction: cross-correlation analysis and adaptive filtering. Algorithms are designed to extract conversations in noisy environment. Such situations can appear in police investigations' materials or multi-speaker environment. Noise can be added intentionally by suspects or not intentionally (e.g. in a car interior). Both of the algorithms are based on recordings from a dual-microphone system. The presented methods use the small differences between recordings. Algorithms were compared taking SNR improvement and better speech understanding into consideration.

### KEY WORDS

source-separation, adaptive filtration, multi-microphone systems

## 1 Introduction

Eavesdropping is one of the most efficient and cheapest sources to provide exhibits of crimes for police and homeland security forces investigations. However, there are many difficulties connected with recordings obtained by listening-in systems. Most of them are caused by random noise or intentionally added disturbances.

One of the solutions to overcome the above problems is based on multi-microphone arrays. Microphone arrays applied in a need of speech enhancement is a well defined field with several methods: beamforming [2], superdirective beamforming [1], postfiltering [5] and phase based filtering [3, 7]. However, all solutions known to authors are focused on solving a problem of a random background noise caused by environment where recording takes place.

It means that they operate on model

$$\begin{aligned} s_{m1,in}(t) &= s_{voice}(t) + n_1(t), \\ s_{m2,in}(t) &= s_{voice}(t - \tau_1) + n_2(t), \end{aligned} \quad (1)$$

where  $s_{voice}(t)$  is a speech signal, and delay  $\tau_1$  is caused by a longer distance to the second microphone. Signals  $n_1(t)$  and  $n_2(t)$  represent microphone and environmental noise respectively. It can be noise of a car machine, traffic noise, disturbances caused by wind or noise of recording system. An important issue is that  $n_1(t)$  and  $n_2(t)$  are uncorrelated.

Our case is significantly different, because the noise was added intentionally by a conversing human to degrade quality of recordings as much as possible. Let us consider the model accommodated to such situation

$$\begin{aligned} s_{m1,in}(t) &= s_{voice}(t) + s_{dist}(t) + n_1(t), \\ s_{m2,in}(t) &= s_{voice}(t - \tau_1) + s_{dist}(t - \tau_2) + n_2(t), \end{aligned} \quad (2)$$

where  $s_{dist}(t)$  is intentionally added disturbance. Delay  $\tau_2$  is not equal to  $\tau_1$  because of differences in distances between microphones and audio signal source.

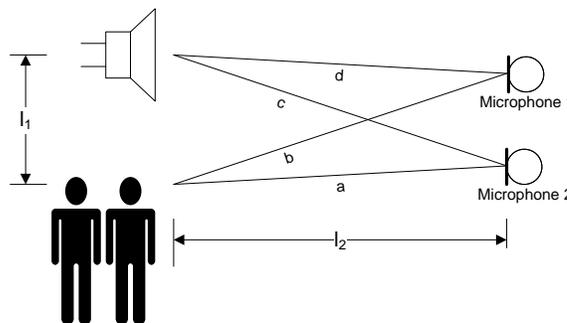


Figure 1. Dual-microphone scenario of listening-in to a conversation where source of a distracting signal, like a radio, was used to hide content of the conversation.

What makes a situation more complicated, the microphones have to be hidden from speakers and in places where it was possible to put a tapping device. This is much different to scenarios typical for information centres or conference rooms. Several efficient methods including a phase-based filtering, which is a form of time-frequency masking (PBTFM) [7] require speaker's position to be known. It is all not possible in our scenario because the speakers are in their homes, cars or jail cells where they can move around and the microphones are listening-in devices. In such case cross-correlation and adaptive filtration algorithms seem to be a good solution.

As the algorithms which we developed are universal, they can be useful in many other, also commercial civil, applications like: noise canceling in automatic speech recognition or hands-free car systems.

The paper is divided as follows. Section 2 presents details on the recording scenario we consider. Section 3 and 4 covers description of cross-correlation and adaptive filtration algorithms respectively. Section 5 provides results of experiments we conducted. The paper is summed up by conclusions.

## 2 Problem description

The problem of separating a conversation from the audio signal is depicted in Figure 1 [9]. The audio signals are acquired by two hidden microphones. There are two speaking persons who use a distracting signal, like music from a radio receiver, 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 verify it, let us assume the sampling frequency 44 100 Hz. Then a time difference between two samples relates to a distance

$$\rho = v\Delta t, \quad (3)$$

where  $v$  is the sound velocity and  $\Delta t$  is the sampling period. For values  $v = 330 \frac{m}{s}$  and  $\Delta t = 23\mu s$  one obtains  $\rho \approx 7.5mm$ . For a real case application, at least a difference of around ten samples between signals from both microphones is needed to proceed. This gives a few centimeters as a necessary difference between the distances. In a very special case, that at once, both  $a \approx b$  and  $c \approx d$  the method would not work. However, this is a very rare scenario in real world situations.

The algorithms described below utilise the differences between distances. The cross-correlation algorithm uses additionally differences in frequency bands: higher for music signal and lower for speech, both detectable by an ear. Figure 2 shows the model of spectral density of a human speech and music (trumpet). It is noticeable that in this case a spectrum of musical instrument is wider than a spectrum of a human speech. What is more, the maximum of spectral energy of music lies in much higher frequencies than the maximum of human voice spectral density.

## 3 Cross-correlation algorithm

The difference between spectral density of speech and music is a basis of cross-correlation algorithm. A block diagram of this algorithm is depicted in Figure 3. Let us assume that  $s_{m1,in}$  and  $s_{m2,in}$  are recordings from Microphone 1 and Microphone 2 (see Figure 1) respectively. Ranges  $d$  and  $c$  are distances from a disturbance source to microphones and ranges  $b$  and  $a$  are distances from talking people to microphones. Our aim is to find  $\tau_2$  from (2), which is a shift in time resulted from a difference between  $d$  and  $c$ . We will use cross-correlation function  $c(\tau)$ , which

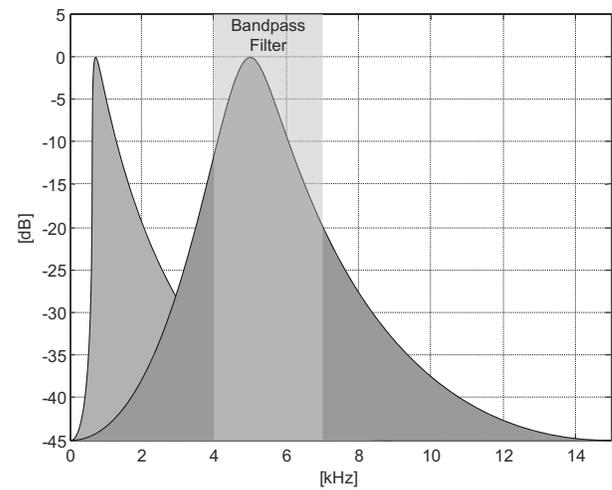


Figure 2. Comparison of spectral density of a speech (light grey) and a trumpet (dark grey). The filter band we were used is marked.

is defined as

$$c(\tau) = \sum_n s_1(n - \tau)s_2(n), \quad (4)$$

to find  $\tau_2$ .

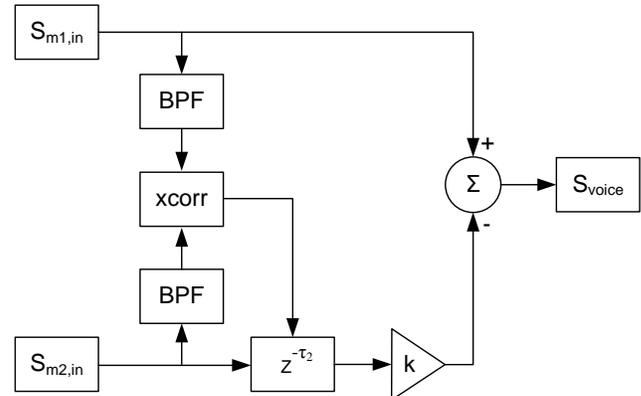


Figure 3. Block diagram of the cross-correlation algorithm of speech extraction.

Time delay  $\tau$  is calculated from cross-correlation taking into account frequency band from 4 to 6.5 kHz which is higher than voice frequency band. Major part of distortion signal energy belongs to this band. It allows to cut-off  $s_{dist}$  and extract  $s_{voice}$ .

Let us assume  $s_1$  and  $s_2$  to be filtered signals from the microphones. The band pass filters are set as it is shown in Figure 2. Such settings allow us to cut off the majority of a speech signal. Then, output signals  $s_{m1,out}$  and  $s_{m2,out}$  contain the music signal mainly, what allows easier calculation of a maximal value of the cross-correlation

(see Figure 4)

$$\tau_2 = \arg \max_{\tau} [\sum_n s_{m1,out}(n - \tau) s_{m2,out}(n)]. \quad (5)$$

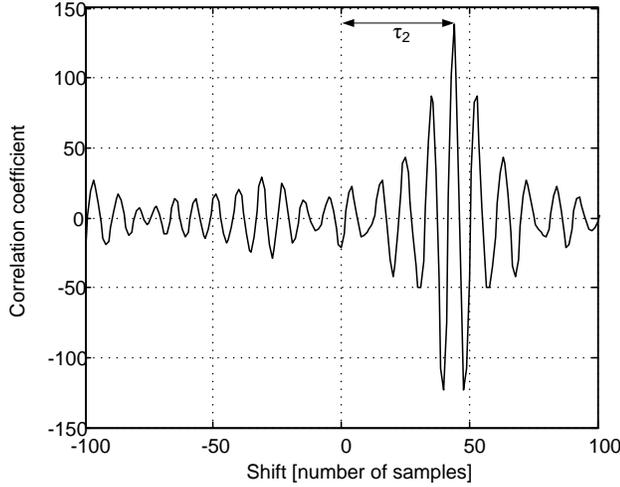


Figure 4. Correlation coefficient as a function of a shift between signals.

The delay determined above is used in a delay block  $z^{-\tau_2}$  (see Figure 3). As a result we get a signal with a compensated impact of the distance. Then the speech signal can be found as

$$s_{voice}(n) = s_{m1,in}(n) - k s_{m2,in}(n - \tau_2), \quad (6)$$

where

$$k = \sqrt{\frac{\sum_n (s_{m1}^{out}(n))^2}{\sum_n (s_{m2}^{out}(n))^2}} \quad (7)$$

is an amplification, which compensates the difference in power of music signal coming from the different distances  $c$  and  $d$ .

In some cases  $\tau_1$  and  $\tau_2$  can be negative. This is why computations are conducted not in real time. It results in additional delay of extracted voice  $s_{voice}$  of amount equal to  $\max(|\tau_1|, |\tau_2|)$ .

## 4 Adaptive filtration algorithm

There are several practical problems which cannot be treated by the algorithm described above. Probably the most important one is existence of reverberations and, generally, various ways of waves propagation for the different frequencies. A filter with a dedicated phase characteristic can be a good solution in this situation. As the location of talking person is not time-invariant, the filter should adapt to the circumstances. Adaptive filtration seems to fulfill all our requirements.

The same input signals which are described by (2) are inputs for the second method we examined. Block diagram of an adaptive filtration algorithm is depicted in Figure 5

[4, 8]. The architecture of the algorithm is different than the previous one. Low pass filters were applied instead of band pass filters. Speech is expected as a final signal so filtering of higher band would unnecessarily make the filtration more complicated. What is more, when wideband signals were filtered, the algorithm did not manage to find proper filter coefficients.

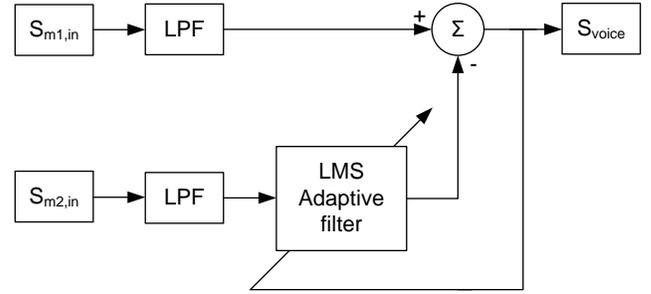


Figure 5. Block diagram of adaptive filtration algorithm of speech extraction.

As the LMS (Least Mean Square) adaptive filter was used, the order for the output  $s_{voice}$  is to be a minimum-mean-squared-error estimate of the signal  $s_{voice}$ . Rodriguez [6] expressed the power of noise in output as (the discrete time index has been omitted for clarity)

$$\begin{aligned} E_n &= E[(s_{voice} - s_{m1,in})^2] \\ &= E[s_{voice}^2] - 2E[s_{m1,in} s_{voice}] + E[s_{m1,in}^2] \\ &= E[s_{voice}^2] - 2E[s_{m1,in}(s_{m1,in} + s_{m2,in} - y)] + E[s_{m1,in}^2] \\ &= E[s_{voice}^2] - E[s_{m1,in}^2] - 2E[s_{m1,in} s_{m2,in}] + 2E[s_{m1,in} y] \\ &= E[s_{voice}^2] - E[s_{m1,in}^2], \end{aligned} \quad (8)$$

where  $y$  is an estimate of the primary noise created by LMS filter. As the signal  $s_{m1,in}$  is unaffected by the adaptive filter, the algorithm sets coefficients of this filter to minimize the total output power  $E[s_{voice}^2]$ .

## 5 Experiments

The results of the implemented algorithms using recordings containing speech disturbed by music were examined. The recordings were produced in our laboratory using two microphones with cardioid beam patterns. Speaking persons and a disturbance source were located in front of microphones. We simulated natural conditions with both speakers and disturbance source simultaneously recorded at the same session.

Both cross-correlation and adaptive filtration algorithms were optimized for the biggest increase of Voice-To-Music Ratio (VMR). To measure VMR of given signals we assumed that the voice signal is unaffected by both algorithms. The signals contain voice disturbed by music. There were segments in which only the music is au-

Algorithm	Increase of VMR [dB]
Cross-correlation	2.0
Adaptive filtering	2.9

Table 1. The improvement results for the described algorithms.

dible due to very low VMR (less than -10). Let us assume that  $P_{m1,voice}$  is average energy of input signal  $s_{m1,in}$  in a range  $(n_1, n_2)$  where voice is disturbed, counted as follows

$$P_{m1,voice} = \frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} s_{m1,in}^2(n). \quad (9)$$

Then  $P_{m1,music}$  would be an average energy of input signal  $s_{m1,in}$  in range  $(n_1, n_2)$  where only music is audible. Therefore let us define VMR of  $s_{m1,in}$  as

$$VMR_{m1} = 10 \log\left(\frac{P_{m1,voice}}{P_{m1,music}}\right). \quad (10)$$

$VMR_{m2}$ ,  $VMR_{out}$  would be the VMRs for  $s_{m2,in}$  and the output signal respectively. Then we count increase  $\Delta VMR$  of VMR as

$$\Delta VMR = VMR_{out} - \overline{(VMR_{m1}, VMR_{m2})}. \quad (11)$$

Table 1 compares results for the both described algorithms by presenting an improvement in VMR.

## 6 Conclusion

The presented methods of signal analysis from two microphones was found successful in recovering conversation. BPFs with a band which is audible for a human ear, but above speech frequencies were found successful in cross-correlation algorithm. Using BFSs with the same settings as in the first method in the adaptive filtration algorithm seems to be unnecessary or even undesirable because of problems with finding proper filter coefficients in frequencies above 4-5 kHz.

The adaptive filtration algorithm gave better results in the examined cases. The main disadvantage of this method is that it is much more complex and computationally demanding what can be important in an end user implementation. Further investigations will be focused on the second method.

The scenario assumptions are that there are very few possible localizations of microphones and they have to be hidden as listening-in devices. The disruptive signal (e.g. music from a radio) is added intentionally by conversing speakers to hide the speech content, along with noise.

## Acknowledgement

This work was supported by MNISW grant OR00001905.

## References

- [1] J. Bitzer, K. U. Simmer, and K. D. Kammeyer. Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement. *Proc. IEEE Int. Conference on Acoustics, Speech, Signal Processing*, 5:2965–2968, 1999.
- [2] G. DeMuth. Frequency domain beamforming techniques. *Proc. IEEE Int. Conference on Acoustics, Speech, Signal Processing*, 2:713–715, 1977.
- [3] D. Halupka, A. S. Rabi, P. Aarabi, and A. Sheikholeslami. Low-power dual-microphone speech enhancement using field programmable gate arrays. *IEEE Transactions on Signal Processing*, 55(7):3526–3535, 2007.
- [4] S. Haykin. *Adaptive Filter Theory*. Prentice-Hall, 1986.
- [5] C. Marro, Y. Mahieux, , and K. U. Simmer. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Trans. Speech, Audio, Signal Process.*, 6:240–259, 1998.
- [6] J. J. Rodriguez, J. S. Lim, and E. Singer. Adaptive noise reduction in aircraft communication system. *Proceedings of IEEE International Conference on ICASSP*, 12:169 – 172, 1987.
- [7] 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*, pages 684–687, 2003.
- [8] S. V. Vaseghi. *Advanced Digital Signal Processing and Noise Reduction*. John Wiley & Sons, 2006.
- [9] M. Ziólko, B. Ziólko, and R. Samborski. Dual-microphone speech extraction from signals with audio background. *Proc. IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2009.