Wiener Filtration for Speech Extraction from the Intentionally Corrupted Signals

Rafał Samborski, Mariusz Ziółko, Bartosz Ziółko, Jakub Galka
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 suggests a speech enhancement approach to an eavesdropping audio system. Speech signal is disturbed by non-stochastic noise. The algorithm is based on recordings from dual-microphone system. The Wiener filter was applied for speech extraction. The algorithm is designed to capture dialogues in noisy environment as well. It uses the small differences between recordings. The differences in speaker and the source of noise localisation together with differences in spectra, enable us to split both signals.

I. INTRODUCTION

Noise causes many difficulties in all of the ASR (Automatic Speech Recognition) systems. Most of them can be assigned to one of the two groups: a) random noise (such as traffic noise, wind noise etc.), b) non-stochastic disturbances (such as music from radio-set or other persons’ conversations). In some cases - like police investigation materials - this kind of disturbances can be intentionally introduced. There are well developed algorithms of denoising the signals corrupted by random noise. Our studies were focused on the second type of disturbances as it is much less examined.

The well known solution of overcoming the above problems is based on multi-microphone arrays [8]. Microphone arrays applied in speech quality enhancement system is a well defined field with several improving methods: beamforming [2], superdirective beamforming [1], postfiltering [4] and phase based filtering [3], [7].

For a basic dual-microphone array working for case of random noise described above, it is assumed that the noise in both microphones \(s_{n1}(t)\) and \(s_{n2}(t)\) are different and uncorrelated. It means that they operate on a model

\[
\begin{align*}
s_{m1}(t) &= s_{\text{voice}}(t) + s_{\text{dist}}(t) + s_{n1}(t), \\
s_{m2}(t) &= s_{\text{voice}}(t - \tau_2) + s_{\text{dist}}(t - \tau_2) + s_{n2}(t),
\end{align*}
\]

(1)

where \(s_{\text{voice}}(t)\) is a speech signal, and delay \(\tau_1\) is caused by a different distances between a speaker and both microphones [9]. Signals \(s_{n1}(t)\) and \(s_{n2}(t)\) represent a microphone and environmental noise. It can be generated by a car machine, street traffic, disturbances caused by wind or noise of a recording system.

The case presented in this paper is significantly different (see Fig. 1), because the considered noise has specific properties. Typically it is added intentionally by talking persons to degrade the quality of recordings as much as possible. As it was mentioned above, such situation can appear in Police investigation materials or multi-speaker environment. Let us consider the model describing such situation

\[
\begin{align*}
s_{m1}(t) &= s_{\text{voice}}(t) + s_{\text{dist}}(t) + n_1(t), \\
s_{m2}(t) &= s_{\text{voice}}(t - \tau_1) + s_{\text{dist}}(t - \tau_2) + n_2(t),
\end{align*}
\]

(2)

where \(s_{\text{dist}}(t)\) is an intentionally added disturbance. The time-shift \(\tau_2\) is not equal to \(\tau_1\) because of differences in distances between microphones and audio signal sources like radio or TV-set [5].

This is much different from scenarios typical for information centres or conference rooms. Position of the speaker is unknown what makes a situation more complicated. Several efficient methods including a phase-based filtering, which is a form of time-frequency masking (PBTFM) [7], require speakers positions to be known. It is not possible in the considered scenario.

The developed algorithm is universal and besides its applications described above, it can be also used in commercial systems like: hands-free car systems and noise cancelling in ASR.

The paper is divided as follows. Section II presents the considered recording scenario. Section III discusses details on the adaptive filtration algorithms. The implementation of the algorithms was described in the Section IV. Section V covers results of the experiments conducted to verify the theoretical considerations. The paper is summed up by conclusions.
II. PROBLEM DESCRIPTION

The problem of separating a conversation from the audio signal is presented in Fig. 1. The audio signals are acquired by two microphones. There are two speaking persons and a disturbing signal, like music from a radio-set, used to block off the understanding of the content of the conversation. In order to proceed with detecting speech signal from the disturbed 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 44100 Hz. Then a time difference between two samples relates to a distance

$$\rho = v \Delta t,$$

where $v$ is the sound velocity and $\Delta t$ is the sampling period. For values $v = 330 \text{ m/s}$ and $\Delta t = 23 \text{ ms}$ one obtains $\rho \approx 7.5 \text{ mm}$. 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 between distances from sound sources and microphones.

Fig. 2 shows the model of the spectral density of the speech and music (trumpet). It is noticeable that spectrum of musical instrument is wider than the spectrum of the speech. What is more, average frequency of music generally covers much higher frequencies than a voice does. This property is sufficient to distinguish the speech and non-speech segments. This observation leads to an important part of the algorithm. Bandpass filters can be used to preserve signals obtained by both microphones. Next, Wiener filter can be tuned to minimise the energy of a difference between filtered signals from the both channels. A tuned filter has a group delay response which compensates the difference in distance to the microphones.

![Fig. 2. Comparison of the model of spectral density of a speech (light grey) and a trumpet (dark grey). The frequency band of the filter used is marked.](image)

III. WIENER FILTERS

The purpose of adaptive filter is to reduce the amount of noise present in signals. The noise components captured by the microphones are filtered and attenuated (see Fig. 3). In case of this scenario, the signal from one of the microphone is treated as the estimation. One seeks the linear filter which response for the first signal would come as similar to the second signal as possible. Adaptive filters are characterised by an assumption that signal and additive noise are linear stochastic processes with known spectral characteristics or known autocorrelation and cross-correlation. It is required that the filter must be realisable, i.e. causal. This requirement can be dropped by adding some delay in signal processing, resulting in a causal solution.

The requirements presented above lead to the solution found by Wiener in the 1940s. The goal of algorithm is to minimize the expected value of the squared error

$$Q = E \{ e^2(n) \},$$

where $e(n)$ is the difference between the observed signal and the estimated signal

$$Q = E \{ (s_{\text{obs}}(n) - s_{\text{est}}(n))^2 \}. $$

The estimated signal is given by

$$s_{\text{est}}(n) = h^T s_{\text{obs},n},$$

where $h \in \mathbb{R}^{N+1}$ is a vector containing the coefficients of the filter and

$$s_{\text{obs},n} = [s_{\text{obs}}(n), s_{\text{obs}}(n-1), \ldots, s_{\text{obs}}(n-N)]^T \in \mathbb{R}^{N+1}.$$

Taking (5) and (6) results in

$$Q = E \{ s_{\text{obs}}^2(n) \} - 2h^T \Phi_{\text{corr}} + h^T \Phi_{\text{obs}} h,$$

where $\Phi_{\text{corr}} = E \{ s_{\text{des}}(n) s_{\text{obs}}(n) \}$ is a vector of cross-correlation and $\Phi_{\text{obs}} = E \{ s_{\text{des}}(n) s_{\text{des}}(n) \}$ is a matrix of autocorrelation.

The vector $h_{\text{opt}}$ (Wiener filter) that minimizes the expression above is given by

$$h_{\text{opt}} = \Phi_{\text{corr}}^\dagger \Phi_{\text{obs}} h,$$

IV. IMPLEMENTATION OF THE ALGORITHM

Fig. 3 depicts the overall architecture of the microphone array speech enhancement system. Algorithm has been implemented in MATLAB and it works as follows:

1) Assumptions:
   - $N$ is the Wiener filter order,
   - matrix $\Phi_{\text{obs},n} = \delta^{-1} I_{N+1}$, where $\delta$ is a small positive coefficient and $I_{N+1}$ is an identity matrix,
   - Wiener filter coefficients $h_{n-1} \in \mathbb{R}^{N+1}$,
   - $h_1 := 0$.

2) For every next sample, the estimated signal and the error signal are given by

$$s_{\text{est}}(n) = h_{n-1}^T s_{\text{obs},n},$$

$$e(n) = s_{\text{des}}(n) - s_{\text{est}}(n).$$
3) The new filter coefficients $h_n$ are calculated using the Sherman-Morrison algorithm [6]:

$$h_n = h_{n-1} + \Phi^{-1}_{obs,n} s_{obs,n} c(n), \quad (11)$$

where

$$\Phi^{-1}_{obs,n} = \Phi^{-1}_{obs,n-1} - \frac{\Phi^{-1}_{obs,n-1} s_{obs,n} s_{obs,n}^T \Phi^{-1}_{obs,n-1}}{1 + s_{obs,n}^T \Phi^{-1}_{obs,n} s_{obs,n}}. \quad (12)$$

The performance criterion of the filter is the minimum mean-square error (MMSE)

$$MMSE = E\left\{ |e(n)|^2 \right\}, \quad (13)$$

which has a physical interpretation of an energy of difference between one of the recorded signals and the second one processed by adaptive filter.

Due to pre-filtering, some frequency bands are attenuated by the Wiener filter more than the others. To preserve the filter from attenuating frequency components outside the 4-7 kHz, the magnitude response of the Wiener filter is flattened by setting up all the zeros of the filter to a unitary circle. The filter phase properties are preserved. A new transmittance of the equalised filter is then given by

$$H(z) = \prod_{n=1}^{N} (z - e^{j\phi_n}), \quad (14)$$

where $\phi_n$ corresponds with the Wiener filter phase components. This all-pass multi-delaying filter is then used to filter the first input signal which is then multiplied by a gain factor

$$k = \sqrt{\frac{\sum_j (s_{m1,j}^2(n))^2}{\sum_j (s_{m1,j}^2(n))^2 + k^2}}, \quad (15)$$

Inverted signal is then added to the second input signal $s_{m2}$ (see Fig. 3) and the output signal is calculated as

$$s_{out,n} = s_{m2,n} - ks_{m1,n} + h_n. \quad (16)$$

Because of the introduced delays, temporal difference between the signals is minimised and speech to noise ratio is increased.

V. Experiments

We examined the method presented above to recover the speech from acoustic signal disturbed by music. Studio recordings made by two omnidirectional, capacitor microphones were used as the inputs. The distance between the microphones varied in different recordings from 10 cm to 1 m. Both sources of distortion and the speakers were distant from the microphone a few meters. The sampling frequency of recordings was 44.1 kHz with the 24-bit resolution. We simulated natural conditions as the laboratory was a live studio with persons speaking naturally.

Two main factors need to be taken into consideration to set the correct Wiener filter of order $N$. The computational complexity of the algorithm is $O(n^2)$, therefore high filter order leads to the long computation time. On the other hand, low filter order can not be sufficient to compensate the difference in distance between microphones. According to (3), every incrementing of the filter order enables compensation of a distance longer than $\rho \approx 7.5$ mm. Taking above into consideration in the conducted experiments $N = 100$ was assumed. Fig. 4 presents the Wiener filter adaptation. Fig. 5 shows values of Wiener filter coefficients after 1 second of adaptation process. The maximum value position corresponds to a physical difference in distances.

The criterion of evaluation of the algorithm was to increase Voice-To-Music Ratio (VMR). It was assumed, that the voice signal is unaffected by both algorithms to be able to measure VMR of given signals. The signals contain voice disturbed by music. There were segments in which only the music is audible due to very low VMR (less then -10). Let us assume that $P_{m1,\text{voice}}$ is an average energy of input signal $s_{m1,in}$ in a time range $(n_1, n_2)$ where voice is disturbed, counted as follows

$$P_{m1,\text{voice}} = \frac{1}{n_2 - n_1 + 1} \sum_{n=n_1}^{n_2} s_{m1,in}^2(n). \quad (17)$$

Then $P_{m1,\text{music}}$ would be an average energy of input signal
There is a clear maximum which corresponds with the difference in distance between the microphones and noise source.

### TABLE I

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Increase of VMR [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncompensated amplitude characteristic</td>
<td>1.8</td>
</tr>
<tr>
<td>Compensated amplitude characteristic</td>
<td>3.6</td>
</tr>
</tbody>
</table>

$s_{m1, in}$ in range $(n_1, n_2)$ where only music is audible. Therefore let us define VMR of $s_{m1, in}$ as

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

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

$$\Delta VMR = VMR_{out} - (VMR_{m1} + VMR_{m2}).$$

Both input and output signals were normalised to have the same energy, so the comparison would be possible.

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

### VI. CONCLUSIONS

Adaptive filters attenuate the noise components of the signal captured by the microphones. Series of speech enhancement experiments were conducted to assess the proposed algorithm. In the experiments on speech in presence of real and simulated noises, we found that the proposed algorithm performs well.

Introduction of the all-pass, magnitude-flattened filter improves speech extraction capability by 2 dB according to the filtration without such solution. However, greater amount of the room acoustical diffusion effect still affects the efficiency of the algorithm.

Presented method allowed to reduce the amount of the intentionally introduced disturbance (music) from the observed signals. Disturbance reduction was strong enough to recover and make the human speech clearly audible and understandable, what was not possible in the case of the pure input signals.

Further work will be conducted to improve the method and increase obtained perceptual and computational quality measures of extracted speech.

### VII. ACKNOWLEDGEMENTS

This work was supported by MNISW grant OR00001905.

### REFERENCES


