Speech Segments Analysis in Reverberant Room and Cocktail Party Interferences Using Microphone Array

Zoran Šarić
Institute of Security, Kraljice Ane bb, 11000 Belgrade, Serbia and Montenegro, sare@yubc.net

Slobodan T. Jovičić1, Srbijanka Turajlić2
School of Electrical Engineering, Kralja Aleksandra 73, 11000 Belgrade, Serbia and Montenegro, jovicic@etf.bg.ac.yu, turajlic@aan.edu.yu

The labeling of the speech segments that belongs to desired speaker or interferences is important for the analysis and processing of the microphone array signals recorded in the room with reverberation and cocktail party interferences. Speech segments labeling proposed in this paper is based on the direct wave detection. The idea is to detect time interval between direct wave and reflections of the walls. The algorithm tests two hypotheses: H0 – there is a direct wave segment, and H1 – there is a mixture of the direct wave, reflections and interferences. Decision function is based on least squares error of the assumed direct wave propagation model. The signals are processed in time domain that provides small time delay of the detection. The proposed detection algorithm is experimentally verified by simulating the room with reverberation. The detection algorithm applied to the ML algorithm for microphone array weights estimation provides better quality of the restored speech signal compared to the ordinary used minimum variance estimation algorithms.

1 Introduction

The need for the separation of the desired speech segments and interference segments in the reverberant room appears when minimum variance (MV) algorithm is applied for the microphone array weights estimation. It is well known the cancellation effect of the desired signal when MV based weights estimation is applied in room with reverberation [1, 2]. The only way to prevent the desired signal cancellation is to separate the segments with pause in desired signal from the interference pause segments and to apply different adaptation algorithms over them [3], [4]. The algorithm for the separation of the desired signal pause segments from the interference pause segments was proposed in [2], [4]. That segments separation algorithm is based on estimation of the signal to interference ratio. The alternative algorithm for segments separation based on the direct wave detection of the desired signal is proposed in this paper. This is accomplished by using the time delay between direct wave and reflection. The decision function for the direct wave detection is the short time mean squares error from assumed direct wave model. When the decision function is over the defined threshold, the decision is direct wave. Contrary, if the decision function is under the threshold, the decision is that there is no direct wave or there are interferences beside direct wave.

Signal detection is realized in the time domain, and its advantage is the short time delay of the detection. The proposed algorithm is tested on simulation of the room with reverberation and applied to the maximum likelihood (ML) algorithm for the desired signal estimation [5]. The results show the better restored signal quality compared to some ordinary adaptive beamforming algorithms.

2 Signals model

The microphone signals are generating by the model

\[ \mathbf{Y}_t = \mathbf{H} \mathbf{s}_t + \mathbf{N}_t = \mathbf{Y}_t. \tag{1} \]

where \( \mathbf{Y}_t \) is \( n \)-column vector of \( n \) microphone signals, \( \mathbf{s}_t \) is \( m \)-column vector of \( m \) acoustic sources. The first source \( s_{0t}, (s_{0t} \in \mathbf{s}_t) \) is desired speaker, while the others are interferences. \( \mathbf{H} \) is \( n \)-by-\( m \) matrix of transfer functions from each source to all microphones. \( \mathbf{N} \) is \( n \)-column vector of additive white noise signals. Components of the vector \( \mathbf{N} \) are mutually uncorrelated, and the its covariance matrix \( \mathbf{R}_N \) is equal to

\[ \mathbf{R}_N = E(\mathbf{NN}^\top) = \sigma^2 \mathbf{I}. \]

Microphone signals are processed by the adaptive algorithm depicted in Figure 1. Time delay of the direct path to the desired speaker is compensated by the
matrix $\Theta$, $\Theta = \text{diag}(q^{-\tau_1}, \ldots, q^{-\tau_n-1})$ by the relation
\[
\Theta Y_t = \Theta H S_t + \Theta N_t = \tilde{H} S_t + \tilde{N}_t - X_t, \tag{2}
\]
where $q^{-\tau_k}$ is time delay operator by $\tau_k$. Matrix $\tilde{H}, \tilde{H} = \Theta H$ includes both transfer matrix $H$ and delay matrix $\Theta$. Vector $\tilde{N}_t, \tilde{N}_t = \Theta N_t$ is equivalent noise.

Vector $\tilde{h}_0$ is the first column of the matrix $\tilde{H}$, and describes transfer from the desired speaker to each component of the signal vector $X_t$. Vector $\tilde{h}_0$ can be decomposed on direct path transfer vector $h_{d0}, h_{d0} = [1, \ldots, 1]'$ and on transfer vector of the reflections $r_{ho}$ by the $\tilde{h}_0 = h_{d0} + r_{ho}$, \tag{3}

Vector $h_{d0}$ includes time delay $\tau_0$ between the first reflection and direct path, and the minimal phase transfer function. Direct wave can be detected by testing following two hypotheses:

**H0**: There is only direct wave signal from the desired speaker described by model $M_0$
\[
X_t = h_{d0}\tilde{s}_0 + \tilde{N}_t \tag{4}
\]

**H1**: There exist reflections or interferences described by model $M_1$
\[
X_t = \tilde{H} S_t + \tilde{N}_t \tag{5}
\]

### 3 Likelihood function

Model $M_0$ can be interpreted as first order regressive model with following meanings:

$X_t$ - $n$-column observation vector,
$h_{d0}$ - known $n$-column vector, $h_{d0} = [1, \ldots, 1]'$
$s_{0t}$ - unobservable parameter numerically equal to desired signal at time $t$.

Superscript * denotes a complex conjugate transpose. By the (4), mean-squares estimate of the desired signal $s_{0t}$ is
\[
\hat{s}_{0t} = \frac{1}{h_{d0}'h_{d0}} h_{d0}'X_t = \frac{1}{n} [1, \ldots, 1]'X_t, \tag{6}
\]

Estimation error vector at time instant $t$ is
\[
E_t = X_t - h_{d0}\hat{s}_{0t}. \tag{7}
\]

Under assumption that noise is Gaussian with variance $\sigma$, logarithm likelihood function $L(X_t | M_0)$ is
\[
L(X_t | M_0) = \log(p(X_t | \tilde{s}_0)) = n \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{\|E_t\|^2}{2\sigma^2} \tag{8}
\]

The likelihood for the $N$ snapshots of the observed signals can be expressed by
\[
L(X_t, \ldots, X_{t+N-1} | M_0) = Nn \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \sum_{k=t}^{t+N-1} \frac{\|E_k\|^2}{2\sigma^2} \tag{9}
\]

### 4 One model detection

In practice the only model $M_0$ is a priori known. As we do not know the positions of the interferences and the transfer functions of the reflections, the parameters of the model $M_1$ is unknown. In that case the solution is to apply the one model based detection [6]. This can be done by surveying of the log-likelihood function and if it is over defined threshold $\lambda$, the decision is hypothesis $H0$. Contrary, the decision is $H1$. The decision role can be defined by
\[
L(X_t, \ldots, X_{t+N-1} | M_0) \geq \lambda \tag{10}
\]

Where $L(X_t, \ldots, X_{t+N-1} | M_0)$ is defined by (9). There is a problem, because the noise variance $\sigma$ is unknown. To overcome this we will apply geometrical method that will be described in the next.

### 5 Geometrical method

The visualization of the signal space in three dimensions is depicted in Figure 2. All components of
the direct wave vector $S_{th}$ are equal and lie on the straight line $s$. The projection of the measured vector $X_i$ on the vector $S_{th}$ is $\hat{S}_{th}$. It represents mean squares estimation of the $S_{th}$ under the assumption that $H0$ is valid.

Let us define decision function $f(X_i)$ as reciprocal value of the square of the $\sin(\theta)$ by

$$f(X_i) = \frac{1}{\sin^2(\theta)} = \sec^2(\theta) = \frac{\|X_i\|^2}{\|E_i\|^2},$$  

(11)

where $\| . \|$ denotes Euclidian norm. Both of the functions $f(X_i)$ and $L(X_i|M_0)$ are strictly decreasing respecting to the mean squares error $\| E_i \|^2$. So, we can substitute $L(X_i|M_0)$ by the $f(X_i)$. The advantage of the function $f(X_i)$ compared to $L(X_i|M_0)$ is that it needn’t the knowledge of the unknown noise variance $\sigma$. Just as $L(X_i|M_0)$, it can be extended to the $N$ snapshots of the observed signals by

$$f(X_1, ..., X_{t+N-1}) = \frac{\sum_{k=N}^{t+N-1} \|X_k\|^2}{\sum_{k=t}^{t+N-1} \|E_k\|^2}.$$  

(12)

The decision role based on $f(X_1, ..., X_{t+N-1})$ is

$$H_0$$

$$f(X_1, ..., X_{t+N-1}) \geq \lambda$$

(13)

6 Optimization

The detection algorithm applied to speech signal has the following drawback. Usually, after pause, the power of the desired speech signal increases slowly, so until the time when the signal power is greatly over the noise floor, reflections from the walls have already come. Practically, there is no time interval where the only direct wave is present. Because of that it is difficult to detect direct wave as it is mixed with reflections. To make detection algorithm robust against reflection waves, we propose two improvements of the algorithm.

Weighted least squares method.

As the power of the speech signal varies in the time, the idea is to enhance the contribution of the time intervals with higher signal-to-noise ratio. Hence, the relation (12) has to be substituted with

$$f_p(X_1, ..., X_{t+N-1}) = \sum_{k=N}^{t+N-1} \alpha_k \|X_k\|^2$$

$$\sum_{k=t}^{t+N-1} \alpha_k \|E_k\|^2,$$

(14)

where $\alpha_1, ..., \alpha_{t+N-1}$ are sequence of the positive numbers used to enhance or reduce the contribution of particular snapshot to decision function $f_p(X_1, ..., X_{t+N-1})$. They are estimated by

$$\alpha_k = \frac{1}{p_{\max}} \min \{p_n(X_k) \leq 1, p_n(X_k) = \|X_k\|^2,$$

$$p_{\max} = \max_k \{p_n(X_k)\}, \quad k = 1, N.$$

Pre-filtering

Each spectral component doesn’t contribute equally to the detection. It is well-known that the most of the speech energy is placed under 1000Hz, and hence, the best signal-to-noise ratio is in that frequency band. Hence, it is useful to apply low-pass filter (0-1000Hz) to reduce the frequency components that corrupt detection because of low signal-to-noise ratio.

7 Application to the ML adaptive beamforming

Maximum likelihood adaptive beamformer has two benefits. The first is maximization of the signal power, and the second is minimization of the interference power [5]. This algorithm calls for the estimation of the two covariance matrices. The first, signal matrix describes room response to the desired signal excitation, while the second interference matrix describes the room response to the interferences excitation. In the real reverberant room where the lot of people talks simultaneously, these two matrices can not be easily estimated from the input snapshots $X$. Nevertheless, this problem can be solved taking into account that in natural speech approximately 20% of time is pauses. Because of that there are time intervals when the only desired speech signal is present as well as there is intervals when the only interferences are present. All we have to do is to apply some detection algorithm to select these intervals. If we select the time intervals with pause in desired speaker, we can estimate interference matrix. On the other side, the signal matrix can be estimated on intervals with pause of the interference. Detection of the pauses in desired speech can be done by the algorithm described in [2, 4]. The intervals with desired signal and with pause in interferences can be selected by the proposed direct wave detection algorithm. After the direct wave detection, there is another time interval where both direct wave and reflections of the desired signal are present. This interval can be used for signal covariance matrix estimation. Unfortunately, the duration of this interval can not be estimated or measured. As the practical solution of the problem, it can be assumed short time interval, long enough for the signal covariance estimation and short enough that the
possibility of the presence of interference is small enough.
Taking the estimates of the signal and interference covariance matrices, the optimal weight vector can be estimated by solving generalized eigenvalue problem as described in [5].

8 Experimental results

The proposed detection algorithm and its application to the ML beamforming algorithm has been verified in a room with reverberation simulated by Allen’s image method [7]. The reverberation time was T_{60}=270ms. There were two sources. The first source s_1 was desired speaker and the second one s_2 was interference (Figure 3). The microphone array consisted of 8 microphone 6cm apart. The sampling rate of the speech signals was 10 kHz. The duration of the each test signal was 10s.

![Figure 3: Experimental setup: simulated room with a reverberation time of 270ms and an 8 microphone array.](image)

The typical time diagrams of the direct wave detection are depicted in Figure 4. Desired speech s_1 is on diagram a), interference speech s_2 is on diagram b) superposition of the s_1 and s_2 on microphone 1 is diagram c), and decision function \( f(X_{1},...,X_{t+1}) \).

Three local maximums in \( f_p(X_{1},...,X_{t+1}) \) point to the instant of the detected direct wave. The amplitude of the local maximum depends on the signal-to-noise ratio as well as the slope of the desired signal power.

![Figure 4: a) desired speaker s_1, b) interference speaker s_2, c) superposition of the s_1 and s_2 on microphone 1, and d) decision function \( f(X_{1},...,X_{t+1}) \).](image)

After the detection of the direct wave there is the interval with both direct wave and reflections. The duration of this interval can not be estimated. Hence, we used fixed time interval of 100ms for signal covariance estimation. Interference covariance matrix is estimated by algorithm presented in [2]. Both covariance matrices are then used for estimation of the weight vector \( W \) by the algorithm presented in [5].

Cepstral distance measure is used for the comparing the following beamforming algorithms:
1) CBF (Conventional Beamformer),
2) Ordinary GSC (Generalized Sidelobe Canceller) with full adaptation,
3) GSC algorithm with weight vector \( W \) estimated on hand labeled pauses in desired speaker [3],
4) GSC algorithm with weight vector \( W \) estimated in ideal case when only interference was present [3],
5) ML algorithm with signal covariance matrix estimated by proposed detection algorithm,
6) ML algorithm with signal covariance and interference covariance matrices were estimated in ideal case when the only signal and only interference was respectively present.

The cepstral distortion measures of restored signal s_1 are shown in Table 1. The restored signal is compared to the room response of the desired signal s_1 on microphone 1. Algorithms are sorted by rising quality of the restored signal s_1.

![Table 1: Cepstral distortion measures of the estimated signal.](image)

<table>
<thead>
<tr>
<th>Estimation method</th>
<th>Cepstral distortion measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. CBF</td>
<td>0.860</td>
</tr>
<tr>
<td>2. GSC with full adaptation</td>
<td>0.758</td>
</tr>
<tr>
<td>3. GSC-hand labeled pauses</td>
<td>0.607</td>
</tr>
<tr>
<td>4. GSC ideal scenario</td>
<td>0.524</td>
</tr>
<tr>
<td>5. ML GSC with proposed direct wave detection</td>
<td>0.496</td>
</tr>
<tr>
<td>6. ML GSC -ideal scenario</td>
<td>0.453</td>
</tr>
</tbody>
</table>
9 Summary

It has been shown that direct wave of the speaker can be successful detected in room with reverberation. The proposed algorithm exploits pauses both in desired speech and interferences. The detection is based on one model that is set up from the position of the desired speaker. The decision function is based on mean squares error from the direct wave model. It has been shown that the proposed detection algorithm can be successfully applied to the ML adaptive beamforming algorithm of the microphone array for estimation of the signal covariance matrix. The experiments show the superiority of the proposed algorithm. Also, potential use of the proposed detection algorithm is acoustical analysis and measurement in room with reverberation.

Acknowledgements

This work was supported in part by the Ministry of Science and Environment Protection of Republic Serbia under Grant number OI-1784.

References


