Exploring Off Time Nature for Speech Enhancement

Meng Yu, and Jack Xin

Department of Mathematics, University of California, Irvine, CA, 92697, USA
myu3@uci.edu, jxin@math.uci.edu

Abstract

It is well known that human conversational speech is “sparse” in time domain, comprising of many “off” time segments. This suggests the utility of the “off” time nature for the task of speech enhancement. We propose an efficient dual-microphone method based on regularized cross-channel cancellation to distinguish the overlapping and single speech segments in the multi-speaker conversational environment. Fortunately, the regularized cancellation results can be reused for speech enhancement along an interference-suppression chain. We present evaluations of the proposed overlapping speech detection and integrated speech enhancement approaches using an IEEE speech database and real room recordings under various acoustic environments, showing promising improvements for speech enhancement by exploiting the off time nature.

Index Terms: Overlapping speech detection, cross-channel cancellation, speech enhancement, $t_i$ optimization.

1. Introduction

Blind speech separation (BSS) aims to recover source signals from their mixtures without detailed knowledge of the mixing process. However, it remains a challenge to retrieve source signals in real-world environments such as in cluttered rooms. Mathematically, the mixing process is convolutive in time and the unknowns are high dimensional. Various efforts have been made to separate convolutive mixtures both in time-domain and frequency-domain [1].

Voice activity detection (VAD) has been widely used in the speech denoising problem [2]. Denoising filters are updated based on noisy speech data, while noise-only segments are used to estimate the statistics of the noise. Motivated by this selectively updating procedure, we ask two questions: is it possible to efficiently enhance individual speech sources from their mixtures based on selected data segments, and which data segments are selected for updating enhancement filters? A simple example suggests a potential answer to the questions. In the case of one speech source and two sensors receiving reverberant source images, blind channel identification (BCI) [3] and blind sparse channel identification (BSCI) [4] identify both channels from source to sensors by cross-channel cancellation. In particular, sparse regularization is exploited for filters resolving the spatial difference between two channels in [4]. This builds an effective acoustic model that not only can resolve solution degeneracies due to the lack of knowledge of the source, but also robust models real acoustic environment. The estimated channel information contributes to the speech dereverberation [4], and also provides a way to suppress the corresponding source speech for the purpose of overlapping speech detection and speech enhancement in this paper.

As observed in [5], conversational speech rarely has a higher than 50% “on” time. Thus, in order to extend the single source model [3, 4] to multi-speaker enhancement model, overlapping speech detection is proposed to distinguish and select those time frames with single speech source by cross-channel cancellation and suppression. Then we achieve to estimate a single speech source after multiple suppression based on the selected data frames. Combining the overlapping speech detection and multi-suppression based speech enhancement, an efficient time-domain convex speech enhancement method is proposed in this paper, assuming that intelligible speech signals contain pauses.

2. Overlapping speech detection (OSD)

Overlapping speech is commonplace at meetings and parties, and poses difficulties for tasks such as source localization, separation and recognition. Hence successful detection of overlapping and non-overlapping speech (single speech) segments is critical to enhance the performance of many speech processing methods. The proposed OSD model adapts two sensors and arbitrary number ($N$) of sources. It carries out detection frame by frame on the observed data streams $x_i$, $i = 1, 2$. The mixing process is $x_i(t) = \sum_{j=1}^N h_{ij} * s_j(t)$, where $t$ is time index, $h_{ij}$ is the room impulse response from source $s_j$ to sensor $i$ and $*$ is linear convolution. Suppose that in a frame $D$, $s_k, k \in \{1, 2, ..., N\}$, is the only active source, it follows from the mixing model that $h_{2k} * x_1(t) - h_{1k} * x_2(t) = 0$ for $t \in D$. The elimination by cross-channel cancellation was known in blind channel identification [3, 4]. Based on the observed data in $D$, a pair of sparse filters $u_i$ ($i = 1, 2$) are sought to minimize the energy of $u_2 * x_1 - u_1 * x_2$ in the region $D$. Ideally, $u_1 \approx h_{1k}$ and $u_2 \approx h_{2k}$. The sparseness of finite impulse response filters regularizes the solution and improves the robustness of the method. Filter sparseness is achieved by $l_1$-norm regularization. The resulting convex optimization problem for $t \in D$ is:

$$
\langle u_1, u_2 \rangle = \arg \min_{u_1, u_2} \frac{1}{2} ||u_2 * x_1 - u_1 * x_2||_2^2 + \frac{\eta}{2} \sum_{i=1}^2 u_i (1 - l_i)^2 + \mu(||u_1||_1 + ||u_2||_1)
$$

(1)

where the second term $\frac{\eta}{2} \sum_{i=1}^2 u_i (1 - l_i)^2$ is to fix scaling and prevent zero (trivial) solution. Denote the length of $D$ by $L_D$ and that of $u_i$ by $l_i$. The frame size of $D$ can be as short as 100 ms. As a result, this spatial difference based method carries out the problem efficiently in terms of the data usage and is different from other OSD methods which rely on high order statistics of data. Since the solution $u_i$ is $l_1$-regularized, the surplus length of it would be 0 while solving (1). In addition, sparseness forces the solution $u_i$ to be able to resolve the major spikes of the channel impulse response filters which comprise the relative time delay. In this sense, it helps to overcome the overfitting problem and be insensitive and robust under reverberant conditions. In
matrix form, convex objective (1) becomes:
\[
u^* = \arg \min_{u} \frac{1}{2} \|Au - f\|^2 + \mu \|u\|_1
\]
(2)
where \(u\) is formed by stacking up \(u_1\) and \(u_2\); vector \(f = (0, 0, \ldots, 0, \eta)^T\) with length \(L_D + 1\), and \((L_D + 1) \times 2L\) matrix \(A\) (\(T\) is transpose) is:
\[
A = \begin{pmatrix}
    x_1(1) & x_1(2) & \cdots & x_1(L_D - 1) & x_1(L_D) & 0 \\
    x_2(1) & x_2(2) & \cdots & x_2(L_D - 1) & x_2(L_D) & 0 \\
    \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
    -x_2(1) & -x_2(2) & \cdots & -x_2(L_D - 1) & -x_2(L_D) & 0 \\
    -x_1(1) & -x_1(2) & \cdots & -x_1(L_D - 1) & -x_1(L_D) & 0 \\
    \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
    -x_1(1) & -x_1(2) & \cdots & -x_1(L_D - 1) & -x_1(L_D) & 0
\end{pmatrix}
\]

with the indices of \(x_1\) and \(x_2\) restricted in \(D\).

For each frame \(D\), we obtain an output signal \(\hat{x}^D = u_1^D \ast x_1 - u_2^D \ast x_2\), where \(u_i^D\) \(i = 1, 2\) are the optimal solutions of (1). Then the whole utterance, \(x\) \((i = 1\ or\ 2)\) and \(\hat{x}^D\) are normalized such that \(L_z\) norm is 1. Finally the energy ratio restricted in \(D\) between normalized \(x\) and \(\hat{x}^D\) is calculated for detection.

\[
R_D = \frac{\|x(\cdot)\|_2^2}{\|\hat{x}^D(\cdot)\|_2^2}
\]
(3)

Mathematically, the cross-channel cancellation doesn’t work if the frame \(D\) contains multiple sources, i.e., \(\|x(\cdot)\|_2^2\) is not much smaller than \(\|\hat{x}^D(\cdot)\|_2^2\). Hence, the significantly large values of this ratio resulting from the good performance of cross-channel cancellation indicate the non-overlapping speech frames. Fig. 1 demonstrates OSD on real room recordings. Two microphones receive the mixtures from four sources on the left panel. The ratio of ground truth (top) tells which frame contains only one active or significantly dominant source, while the ratio of detection (bottom) is the output of the proposed method, which agrees qualitatively with the ground truth in Fig. 1.

In order to select the single source frames, we use an automatic threshold setting method. We denote \(r_{\text{mean}}\) as the mean of the ratio metric in a window of \(2T_{\text{max}}\) time frames around the \(n\)-th frame:
\[
r_{\text{mean}} = \frac{1}{2T_{\text{max}} + 1} \sum_{i=n-T_{\text{max}}}^{n+T_{\text{max}}} R_i,
\]
(4)

with \(n - T_{\text{max}} \leq i \leq n + T_{\text{max}}\). The threshold at \(n\)-th frame is thereby set to
\[
th_{\alpha} = \max(r_n, \alpha r_{\text{mean}}).
\]
(5)

where \(r_0\) is a base threshold uniformly for all the frames and \(\alpha\) is a factor of significance. The \(n\)-th frame is identified as the single source frame if the detection ratio \(R_n\) is above \(\theta_{\alpha}\).

3. Speech enhancement model & analysis

Speech enhancement for two sensors and two sources case is identical with the proposed OSD method. At the selected single speech frame \(D\), the output data stream \(\hat{x}^D = u_1^D \ast x_1 - u_2^D \ast x_2\) cancels out the interference which is active in frame \(D\). Thus, the enhancement is achieved during OSD stage.

For generic model, let us consider the enhancement for the setup with 3 sources and 3 mixtures, where each mixture is the sum of sources coming from different channels as
\[
x_1 = h_{11} \ast s_1 + h_{12} \ast s_2 + h_{13} \ast s_3,
\]
\[
x_2 = h_{21} \ast s_1 + h_{22} \ast s_2 + h_{23} \ast s_3.
\]
\[
x_3 = h_{31} \ast s_1 + h_{32} \ast s_2 + h_{33} \ast s_3.
\]
(6)

Suppose frame \(D\) is identified as \(s_1\)-only frame, while frame \(D_2\) contains only source \(s_2\). In order to enhance the source \(s_1\), two of the three mixtures are chosen for cross-channel cancellation, e.g., \(x_1\) and \(x_2\), with \(u_1^D\), \(i = 1, 2\) as solved cancellation filters. A new estimated signal \(\hat{x}_1^D\) is obtained
\[
x_1^D = u_2^D \ast x_1 - u_1^D \ast x_2
\]
\[
\approx (h_{33} \ast h_{11} - h_{13} \ast h_{31}) \ast s_1
\]
\[
+ (h_{23} \ast h_{12} - h_{13} \ast h_{22}) \ast s_2.
\]
(7)

Similarly, by choosing another pair of mixtures, e.g., \(x_1\) and \(x_3\), with cancellation filters \(u_1^D\), \(i = 1, 3\) accordingly \((u_1^D\) in (8) and that in (7) are not necessarily the same in the computation), we obtain another estimated signal
\[
x_2^D = u_3^D \ast x_1 - u_1^D \ast x_3
\]
\[
\approx (h_{33} \ast h_{11} - h_{13} \ast h_{31}) \ast s_1
\]
\[
+ (h_{23} \ast h_{12} - h_{13} \ast h_{22}) \ast s_2.
\]
(8)

Ideally, \(x_1^D\) and \(x_2^D\) are the mixed data streams of \(s_1\) and \(s_2\) Restricted in \(D_2\), cross-channel cancellation (2) between \(x_1^D\) and \(x_2^D\) solves a pair of sparse filters \(u_1^D\) and \(u_2^D\), which formulate the estimate of source speech \(s_1\),
\[
\hat{s}_1 = u_2^D \ast x_1^D - u_1^D \ast x_2^D
\]
\[
\approx (h_{33} \ast h_{12} - h_{13} \ast h_{32}) \ast (h_{33} \ast h_{11} - h_{13} \ast h_{31}) - (h_{23} \ast h_{12} - h_{13} \ast h_{22}) \ast (h_{33} \ast h_{11} - h_{13} \ast h_{31}) \ast s_1
\]
(9)

Fig. 2 shows the diagram of the overlapping speech detection based speech enhancement algorithm. The input data are \(M\) mixtures of the same number of sources. Arbitrary two of them are selected for overlapping speech detection. The detection is carried out frame by frame. For each identified single source frame, we label it according to the spatial difference between the resolved sparse filters \(u_1\) and \(u_2\) (i.e., direction of arrival). To enhance source \(s_k\), the detection stops once \(M - 1\) frames \(D_i, i = 1, 2, \ldots, M - 1\) are found, corresponding to unrepeatable \(M - 1\) sources but \(s_k\). On the following stages along the chain, each stage receives input data from up-flow and transmits the output data to the down-flow. The single active source in each frame \(D_i\) is suppressed in the output data streams. The initial input data towards frame \(D_1\) is the mixtures \(x_{1, i}, i = 1, 2, \ldots, M\). \(M - 1 + 1\) input data streams are paired for cross-channel cancellation in frame \(D_1\), generating \(M - 1\) output data streams. An available example in Fig. 2 demonstrates that one signal is paired with the rest for cross-channel cancellation at each stage.
4. Experimental evaluation

4.1. Evaluation of OSD

The parameters for the proposed method are set as $\mu = 10^{-3}$, $\eta = 1$ and $\lambda = 2\mu$. The detection is conducted frame by frame with 50% overlapping and the frame size is set as 200 ms. The base threshold $r_0 = 10$, $T_{\text{max}} = 3$, and the factor $\alpha = 2$. The filter length $l$ is 128, 256, 512, 1024 and 1024 taps under five reverberant conditions ($T_{\text{do}}$ of 0 ms (anechoic), 200 ms, 400 ms, 600 ms and 800 ms) respectively. The performance of the proposed method is evaluated under various reverberant conditions by calculating the Recall Rate $R_c$, the Precision Rate $R_p$, the False Alarm Rate $R_f$, and the F-measure. As the F-measure is the harmonic mean of the precision and recall, this score can be taken as the most important measure of the test.

Room impulse responses are created using the Roomsim simulator [6] to synthesize mixed speech signals by IEEE sentences [7], each with 3 s duration and 16k Hz sampling frequency. For two sources case, the azimuth angles of the two sources are fixed at $135^\circ$ and $10^\circ$. For four sources case, sources from the directions $120^\circ$, $135^\circ$, $150^\circ$ and $165^\circ$ synthesize two mixtures. At each reverberation level ($T_{\text{do}}$ of 0 ms (anechoic), 200 ms, 400 ms, 600 ms and 800 ms), a total number of 400 pairs (quadruples) of source signals are randomly selected from the database, and each of them are used to synthesize two mixtures. For each utterance of mixture, about 25% frames are non-overlapping speech frames. From the result shown in Fig. 3 & 4, it can be seen that though reverberation degrades the detection accuracy when it goes up from $T_{\text{do}} = 0$ ms to 800 ms, the proposed OSD method achieves high accuracy for both cases. In the anechoic environment, the correct detection rate is nearly 90%, while this rate is more than 70% under moderate reverberant conditions and even around 50% or more for highly reverberant conditions.

Table 1 evaluates the detection quality affected by the different sizes of non-overlapping segments in the utterances. If the duration of non-overlapping speech segment is extremely small, i.e. around 100 ms, it results in a lower correct detection rate, nevertheless, still around 50%.

Table 1: Non-overlapping speech detection on the room recordings under various sizes of non-overlapping segments

<table>
<thead>
<tr>
<th>Non-overlapping</th>
<th>$R_c$</th>
<th>$R_p$</th>
<th>$R_f$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 ms</td>
<td>0.7188</td>
<td>0.9127</td>
<td>0.0491</td>
<td>0.8042</td>
</tr>
<tr>
<td>400 ms</td>
<td>0.8015</td>
<td>0.8203</td>
<td>0.1192</td>
<td>0.8108</td>
</tr>
<tr>
<td>300 ms</td>
<td>0.7190</td>
<td>0.9063</td>
<td>0.0377</td>
<td>0.8018</td>
</tr>
<tr>
<td>200 ms</td>
<td>0.7424</td>
<td>0.7879</td>
<td>0.1498</td>
<td>0.7645</td>
</tr>
<tr>
<td>100 ms</td>
<td>0.4293</td>
<td>0.4731</td>
<td>0.2382</td>
<td>0.4501</td>
</tr>
</tbody>
</table>

4.2. Evaluation of integrated speech enhancement (OSD-SE)

The comparison of a list of existing speech enhancement methods [1] is shown in Fig. 5 & 7 in terms of average increase in SIR (signal to interference ratio), and SDR (signal to distortion ratio). These methods include Parra’s decorrelation based method, spatial-temporal fast ICA (STFICA), scaled natural gradient method (SNGTD), independent vector analysis method (IVA) and DUET method [1]. Test is based on the synthetic mixtures of two speech signals under various reverberant environments ($T_{\text{do}} = 200$ ms, 400 ms, 600 ms, and 800 ms) used in the evaluation of OSD method. One of the source signals is considered as the interference. The SIR before processing is about -5 dB. Fig. 5 & 7 indicates that the proposed OSD-SE achieves promising separation quality in objective measures. 22 dB of SIR improvement is achieved under low reverberation condition.
Figure 5: Comparison of average improvement in SIR among speech enhancement methods

Figure 6: Configuration and parameters of the room recording. For the two sensors and two sources case, sources come from speaker S1 and S2, and Mic2 and Mic3 are turned on; while for the case of three sensors and three sources, all the speakers and microphones in the room are included.

(T60 = 200 ms), while 9 dB is achieved under highly reverberant environment (T60 = 800 ms). Room recordings are used to evaluate and compare the above speech enhancement methods by the Perceptual Evaluation of Speech Quality (PESQ) [8] and computational time, shown in Table 2.

5. Conclusion

We present an overlapping speech detection method and its associated speech enhancement approach. Relying on overlapping speech detection, speech “off” time nature is explored and contributes to the sparse regularized cross-channel cancellation. This dual microphone based OSD method shows promising results in various acoustic environments. The integrated speech enhancement approach inherits cross-channel estimates from OSD stage and enhances the desired speech source by multiple suppression of interferences. Sparseness constraint imposed on the solution filters renders the method robust in reverberant environments, and the experimental results demonstrate that the “off” time nature based OSD-SE approach holds the potential to robustly solve the speech enhancement problem.

6. References


Table 2: Average PESQ of BSS methods on real recording mixtures. PRE PESQ is the average PESQ of the mixture data.

<table>
<thead>
<tr>
<th></th>
<th>2 sources (time[s])</th>
<th>3 sources (time[s])</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRE PESQ</td>
<td>1.37</td>
<td>1.00</td>
</tr>
<tr>
<td>Parra</td>
<td>1.57 (7.9)</td>
<td>1.44 (16.0)</td>
</tr>
<tr>
<td>STFICA</td>
<td>1.90 (2.1)</td>
<td>1.70 (3.3)</td>
</tr>
<tr>
<td>SNGTD</td>
<td>2.07 (120)</td>
<td>1.88 (265)</td>
</tr>
<tr>
<td>IVA</td>
<td>2.35 (49.0)</td>
<td>2.02 (52.2)</td>
</tr>
<tr>
<td>DUET</td>
<td>2.36 (2.2)</td>
<td>2.00 (4.3)</td>
</tr>
<tr>
<td>OSD-SE</td>
<td>2.53 (4.7)</td>
<td>2.24 (12.9)</td>
</tr>
</tbody>
</table>


