Improving Channel Decorrelation for Multi-Channel Target Speech Extraction

Jiangyu Han\textsuperscript{1,2,*}, Wei Rao\textsuperscript{2}, Yannan Wang\textsuperscript{2}, Yanhua Long\textsuperscript{1\dagger}

\textsuperscript{1}Shanghai Engineering Research Center of Intelligent Education and Bigdata, Shanghai Normal University, Shanghai, China
\textsuperscript{2}Tencent Ethereal Audio Lab, Tencent Corporation, Shenzhen, China

\texttt{jyhan03@163.com, \{ellenwrao,yannanwang\}@tencent.com, yanhua@shnu.edu.cn}

Abstract

Target speech extraction (TSE) aims to extract only a target signal from the mixed speech [1]. Generally, most TSE approaches require additional target speaker clues to drive the network towards extracting the speech of that speaker. Many previous works have focused on the TSE tasks [2–6]. Although these works have achieved great success on close-talk target speech extraction tasks, the performance of far-field speech extraction is still far from satisfactory due to the reverberation, noise, etc. In that case, the additional multi-channel spatial information usually can be used to improve the separation ability.

Many contributions have been made to exploit the spatial information between multi-channel recordings for TSE. For example, the direction aware SpeakerBeam [7] combines an attention mechanism with beamforming to enhance the signal of the target direction; the neural spatial filter [8] uses the directional information of the target speaker to extract the corresponding speech; the time-domain SpeakerBeam (TD-SpeakerBeam) [9] incorporates the inter-microphone phase difference (IPD) [10] as additional input features to further improve the speaker discrimination capability. All of these approaches showed promising results, which indicates that the multi-channel information can provide an alternative guider to discriminate the target speaker better. Actually, one of the keys of TSE is still speech separation that refers to extract all overlapping speech sources from the mixed speech [11]. In order to further improve the separation ability, many strategies for exploiting the multi-channel information have been recently proposed, such as normalized crosscorrelation (NCC) [12], transform-average-concatenate (TAC) [13], and inter-channel convolution difference (ICD) [14], etc. Therefore, how to effectively exploit the multi-channel spatial information for TSE is crucial.

Considering that the mixed speech in an $N$-dimensional embedding space can be decomposed into $N$ representations at the dimensional-level. Given multi-channel input mixtures, each dimensional-level representation among different microphone channels not only has high correlations, but also has some differentiations, both of these correlations and differentiations characterize the inter and intra speaker properties in the mixture. Our previous work in [15] proposed a channel decorrelation (CD) mechanism with parallel encoder target speaker adaptation to leverage these information for improving TSE systems. This decorrelation is performed on each dimension of all the multi-channel encoder representations of input mixtures, it is used to extract the inter-channel differential spatial information to learn difference between individual source signals of input mixture. Results in [15] have already shown that our original CD significantly improved the TSE performance over IPD features, however, the performance gains over the parallel encoder architecture [16] are still limited.

Therefore, in this study, we revisit the principle of our original channel decorrelation algorithm, two improvements are proposed to replace the original softmax-based inter-channel similarity calculation, one is using an unrolled probability and the other is using a normalized cosine-based similarity. Both of them are performed at the dimensional-level of encoder representations, and they aim to expand the dynamic range of the inter-channel differential probability to capture the fine-grained representation for a better spatial information utilization. Moreover, new combination strategies of the improved CD and target speaker adaptation of parallel encoder outputs are also investigated. All of experiments are performed on the publicly available multi-channel reverberant WSJ0 2-mix dataset. Results show that both of our improved channel decorrelation and the new adaptation strategy are very effective to improve the target speech extraction.

1. Introduction

1.1. Conv-TasNet

The Conv-TasNet is a time-domain speech separation technique proposed in [17]. It has become the mainstream speech separation approach because of its competitive performance over most of time-frequency domain speech separation algorithms. This separation structure has attracted widespread attention and

\textsuperscript{*}work done during internship at Tencent Ethereal Audio Lab

\textsuperscript{†}corresponding author
further improved in many recent works [18–21]. Conv-TasNet consists of three parts: an encoder (1d convolution layer), a mask estimator (several convolution blocks), and a decoder (1d convolution layer). The waveform mixture is first encoded by the encoder and then is fed into the temporal convolutional network (TCN) [22] based mask estimator to estimate a multiplicative masking function for each source. Finally, the source waveforms are then reconstructed by transforming the masked encoder representations using the decoder. More details of Conv-TasNet can be found in [17].

2.2. TD-SpeakerBeam

TD-SpeakerBeam is a Conv-TasNet based target speech extraction approach that has been recently proposed in [9]. As shown in Fig.1 (a), it introduces an auxiliary network with an encoder and a convolution block to transform a pre-saved target speaker enrollment utterance of the target speaker, and the target speaker of the first channel, the extracted target speech waveform, the channel decorrelation is proposed to construct such a labeling can be seen as one type of fine-grained representation that corresponds to different local traits of the whole input mixture. Furthermore, by concatenating the IPD features after the adaptation layer, the performance of TD-SpeakerBeam can be further improved.

In Fig.1 (a), y1, x∗, a∗, and e∗ are the mixture waveform of the first channel, the extracted target speech waveform, the enrollment utterance of the target speaker, and the target speaker embedding vector respectively. We use this IPD configuration as one of our baseline.

3. Channel Decorrelation

3.1. Original CD

Assuming there is a N-dimensional embedding space, each channel of the multi-channel mixture waveform can be encoded into a N × T embedding matrix, where T is the length of that matrix. Each dimensional-level vector of the mixture embedding can be seen as one type of fine-grained representation that corresponds to different local traits of the whole input mixture.

Our channel decorrelation is proposed to construct such a latent embedding space to learn dimension-level discriminative feature representations between the encoded representations of multi-channel input mixtures, by taking the advantage of spatial configuration of multi-channel microphone-arrays. Taking two channels for example, the CD process is shown in Fig.1 (b). It accepts two encoded mixture representations W1 and W2, and then outputs the differential spatial information Wcd between two channels. The specific calculation of Wcd is as follows:

First, compute the cosine correlation between W1 and W2 in each corresponding dimension (row). Note that W1 and W2 are two matrices, i.e.,

\[
W_i = [w_{i1}, w_{i2}, ..., w_{iN}]^T, \quad i = 1, 2
\]

where \( W_i \in \mathbb{R}^{N \times T} \) is the input of the \( i \)th channel of CD mechanism, \( N \) is the output dimension of convolutional encoder, \( T \) is the length of convolutional encoder output. \( w_{ij} \in \mathbb{R}^{N \times T}, j = 1, 2, ..., N \), is the \( j \)th dimension vector of the \( i \)th channel. \( T \) is the operation of transpose.

The cosine correlation of \( j \)th dimensional vector between the first channel and the second channel is calculated as,

\[
\phi_{1,2}^j = \frac{\langle w_{1j}, w_{2j} \rangle}{\|w_{1j}\|_2 \|w_{2j}\|_2}, \quad j = 1, 2, ..., N
\]

where \( \langle \cdot, \cdot \rangle \) is the inner product of two vectors, \( \| \cdot \|_2 \) represents the Euclidean norm. The vectors involved in the operation are normalized to zero mean prior to the calculation.

Then, calculate the cosine correlation between the vectors of each dimension of W1 and W2 in turn, and concatenate them to get a similarity vector \( \phi_{1,2} \).

\[
\phi_{1,2} = [\phi_{1,2}^1, \phi_{1,2}^2, ..., \phi_{1,2}^N]^T
\]

where \( \phi_{1,2} \in \mathbb{R}^{N \times 1} \) represents the similarity of two encoded mixture representations in each dimension in a latent space.

Next, introduce an auxiliary vector \( a \) with the same size as \( \phi_{1,2} \) and values are all 1, i.e.,

\[
a = [1, 1, ..., 1]^T
\]

where \( a \in \mathbb{R}^{N \times 1} \) can be regarded as the cosine correlation between first channel \( W_1 \) itself, which represents the reference vector of \( \phi_{1,2} \). Then, a softmax operation is used to calculate the probability between each corresponding dimensional element in \( \phi_{1,2} \) and \( a \) as \( p_{1,2}^j \) to get a similarity probability vector \( p_{1,2} \in \mathbb{R}^{N \times 1} \) as,

\[
p_{1,2}^j = \frac{e^{\phi_{1,2}^j}}{e^{\phi_{1,2}^1} + e^{\phi_{1,2}^2} + ... + e^{\phi_{1,2}^N}}, \quad j = 1, 2, ..., N
\]

Next, subtract \( p_{1,2} \) from \( a \) to get a vector \( s_{1,2} \in \mathbb{R}^{N \times 1} \) that represents differentiated scores between channels, and repeat \( s_{1,2} \) to the same size as \( W_2 \) to get the differentiated score matrix \( S_{1,2} \in \mathbb{R}^{N \times T} \), i.e.,

\[
s_{1,2} = a - p_{1,2}
\]

\[
S_{1,2} = [s_{1,2}, s_{1,2}, ..., s_{1,2}]
\]
Finally, the differentiated spatial information $W_{cd}$ between channels can be obtained by multiplying $W_{2}$ by $S_{1,2}$ as

$$W_{cd} = W_{2} \circ S_{1,2}$$  \hspace{1cm} (7)$$

where $W_{cd} \in \mathbb{R}^{N \times T}$, $\circ$ denotes element-wise multiplication. In addition, to better guide the network towards extracting speech of target speaker, we also perform the target speaker scaling adaptation on $W_{cd}$ to exploit the target speaker-dependent spatial information, as shown in Fig.1 (b).

### 3.2. Unrolled CD

Note that in Equation (5), the softmax-based inter-channel similarity $p_{1,2}^{j}$ ranges from $1/e$ to $1$ or 0.12 to 0.5, which results in a $s_{1,2} \in [0.5, 0.88]$ differentiated score range. This means that the original CD tends to assign a bigger but coarse weight to the inter-channel differential information than the correlations. This is incompatible with our motivation. Intuitively, ideal inter-channel differential information of the same utterance from different microphone channels should be less but fine and discriminative enough.

In order to alleviate such a problem, we modify the Equation (5) as

$$p_{1,2}^{j} = \frac{2e^{\phi_{1,2}^{j}}}{e + e^{\phi_{1,2}^{j}}}, j = 1, 2, ..., N$$  \hspace{1cm} (8)$$

By doing so, we can achieve a $[0, 1]$ inter-channel similarity range and the resulting differentiated score range becomes $[0, 0.76]$. We call this modification as “unrolled CD”. By expanding the dynamic range of $s_{1,2}$ distribution, the unrolled CD may result in more fine and discriminative inter-channel differential information.

### 3.3. Cosine-based CD

Another method to address the limitation of original CD that mentioned in section 3.2 is to directly normalize the cosine-based inter-channel similarity to $[0,1]$ using the conventional cosine probability transformation $(1 + \cos \phi_{1,2})/2$. In this way, the resulting differentiated score is defined as

$$s_{1,2} = \frac{a - \phi_{1,2}}{2}$$  \hspace{1cm} (9)$$

And the dynamic range of $s_{1,2}$ becomes $[0, 1]$. We call this modification as “cosine-based CD”.

### 3.4. CD with Parallel Encoder Speaker Adaptation

Besides the improvements of the original CD algorithm, we also investigate new combination strategies of the CD-based spatial information and target speaker adaptation of parallel encoder outputs. Subfigure (a) and (b) in Fig.2 demonstrate two new different combination strategies, in both of which the output of CD only plays a role of effective inter-channel discriminative spatial information as the IPD used in TD-SpeakerBeam. These two proposals are motivated by the speculation that the original CD tends to assign a bigger but coarse weight to the inter-channel differential information.

![Figure 2: Different combination strategies of the CD-based spatial information and target speaker adaptation of parallel encoder outputs.](image)

### 4. Experiments

#### 4.1. Dataset

Our experiments are performed on the multi-channel reverberant WSJ0-2mix corpus [24]. Multi-channel recordings are generated by convolving clean speech signals with room impulse responses simulated with the image method for reverberation time of up to about 600ms [9]. The dataset consists of 8 channel recordings, but to have a fair comparison with the state-of-the-art baselines, we also use only two channels in experiments.

We use the same way as in [25] to generate adaptation utterances of the target speaker. The utterance is anechoic and is selected randomly that different from the one in the mixture. The size of training, validation, and test sets are 20k, 5k, and 3k utterances, respectively.

#### 4.2. Configurations

All experiments are performed based on our previously released open source of the original CD-based TSE in [15, 26]. We still use the same hyper-parameters as our original CD [15]. However, it’s worth noting that in this study, we only use the scale-invariant signal-to-distortion ratio (SI-SDR) [27] as the training objective function, instead of the multi-task loss with a cross-entropy speaker identification loss used in our previous work of [15]. For evaluating the performance, besides the signal-to-distortion ratio (SDR) of BSSeval [28] and SI-SDR, we also report the PESQ [29] and STOI [30] to measure the speech quality and intelligibility.

#### 4.3. Results in SDR/SI-SDR

Table 1 shows the performance comparison among different frameworks in SDR/SI-SDR for female-female (FF), male-male (MM), female-male (FM) and the whole (Avg) conditions of the reverberated WSJ0-2mix evaluation sets. The first ten systems, from “TSB [9]” to “CD-Old-IPD”, are all mentioned in our previous work of [15], but their performances are slightly worse (absolute $<0.2\text{dB SDR}$) than the numbers we reported in [15]. Because they are all only trained using the SI-SDR instead of the multi-task loss.

“TSB [9], TSB-IPD [9]” are the state-of-the-art TD-SpeakerBeam (TSB) without and with IPD features. Because the source code of TSB is not publicly available, we implemented this algorithm in [26] and the reproduced results shown in “TSB (our), TSB-IPD (our)” are slightly better than the ones in [9]. Instead of using IPD to exploit the multi-channel spatial information, the parallel encoder (Para-Enc) [16] was proposed.
### Table 1: SDR/SI-SDR (dB) performance of different target speech extraction systems. “SA” represents the speaker adaptation is also performed on the parallel encoder output. “FF”, “MM”, “FM” and “Avg” represent the female-female, male-male, female-male and average conditions. All the systems are trained only using the SI-SDR loss.

<table>
<thead>
<tr>
<th>System</th>
<th>SA</th>
<th>FF</th>
<th>MM</th>
<th>FM</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSB-IPD [9]</td>
<td>9.43 / 8.84</td>
<td>10.02 / 9.52</td>
<td>12.54 / 12.06</td>
<td>11.26 / 10.76</td>
<td></td>
</tr>
<tr>
<td>TSB-IPD (our)</td>
<td>10.01 / 9.46</td>
<td>10.51 / 10.02</td>
<td>12.80 / 12.31</td>
<td>11.65 / 11.15</td>
<td></td>
</tr>
<tr>
<td>CD-Unrolled</td>
<td>11.31 / 10.68</td>
<td>11.91 / 11.41</td>
<td>13.31 / 12.79</td>
<td>12.56 / 12.02</td>
<td></td>
</tr>
<tr>
<td>CD-Old</td>
<td>11.16 / 10.53</td>
<td>11.67 / 11.10</td>
<td>13.20 / 12.69</td>
<td>12.41 / 11.85</td>
<td></td>
</tr>
<tr>
<td>CD-Old unrolled</td>
<td>11.53 / 10.92</td>
<td>12.02 / 11.52</td>
<td>13.60 / 13.09</td>
<td>12.78 / 12.26</td>
<td></td>
</tr>
<tr>
<td>CD-Old-IPD</td>
<td>11.20 / 10.58</td>
<td>11.79 / 11.27</td>
<td>13.39 / 12.88</td>
<td>12.55 / 12.01</td>
<td></td>
</tr>
<tr>
<td>CD-Unrolled</td>
<td>11.41 / 10.78</td>
<td>12.10 / 11.60</td>
<td>13.64 / 13.14</td>
<td>12.81 / 12.28</td>
<td></td>
</tr>
<tr>
<td>CD-Cosine</td>
<td>11.33 / 10.70</td>
<td>12.27 / 11.77</td>
<td>13.78 / 13.28</td>
<td>12.92 / 12.39</td>
<td></td>
</tr>
<tr>
<td>CD-Cosine-Tied</td>
<td>11.53 / 10.89</td>
<td>12.08 / 11.56</td>
<td>13.57 / 13.06</td>
<td>12.78 / 12.25</td>
<td></td>
</tr>
<tr>
<td>CD-Para-a</td>
<td>11.71 / 11.10</td>
<td>12.28 / 11.76</td>
<td>13.80 / 13.28</td>
<td>13.00 / 12.46</td>
<td></td>
</tr>
<tr>
<td>CD-Para-b</td>
<td>11.41 / 10.74</td>
<td>12.15 / 11.64</td>
<td>13.68 / 13.17</td>
<td>12.84 / 12.30</td>
<td></td>
</tr>
</tbody>
</table>

### 4.4 Results in PESQ/STOI

<table>
<thead>
<tr>
<th>System</th>
<th>SA</th>
<th>PESQ</th>
<th>STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixture</td>
<td>-</td>
<td>1.964</td>
<td>0.742</td>
</tr>
<tr>
<td>Para-Enc</td>
<td>-</td>
<td>2.987</td>
<td>0.910</td>
</tr>
<tr>
<td>CD-Unrolled</td>
<td>✓</td>
<td>3.047</td>
<td>0.915</td>
</tr>
<tr>
<td>CD-Para-a</td>
<td>✓</td>
<td>3.143</td>
<td>0.918</td>
</tr>
<tr>
<td>CD-Para-b</td>
<td>✓</td>
<td>3.142</td>
<td>0.920</td>
</tr>
</tbody>
</table>

Table 2 shows the PESQ and STOI performance comparison among different systems. Compared with “Para-Enc (w/o SA)”, the “CD-Unrolled (w/o SA)” achieves 0.120 and 0.007 improvements in PESQ and STOI respectively. Furthermore, we also see that the additional speaker adaptation performed on the parallel encoder output is also beneficial for improving the speech quality and intelligibility.

### 5. Conclusion

In this study, two methods are proposed to improve our original channel decorrelation algorithm. By expanding the dynamic range of decorrelation distribution, the improved algorithm can capture better fine-grained inter-channel differential information. Furthermore, new combination strategies of the improved decorrelation and additional target speaker adaptation of encoder outputs are also investigated. Extensive experimental results demonstrate that both the improvements on the original channel decorrelation and target speaker adaptation are very effective to build a much better target speech extraction system. Our future work will focus on generalizing the proposed algorithms to other related tasks.

### 6. Acknowledgements

The work is supported by the National Natural Science Foundation of China (Grant No.62071302).
7. References


[26] https://github.com/lyhan03/channel decorrelation.


