A comparison of simultaneous 3-channel blind source separation to selective separation on channel pairs using 2-channel BSS

Erik Visser, Kwokleung Chan, Stanley Kim, Te-Won Lee

SoftMax Inc
4180 La Jolla Village Drive Suite # 455
La Jolla CA 92037
(evisser,kwchan,tlee)@softmax.com

Abstract

A number of real-life speech applications using BSS have been reported for two channel applications but only a few have been reported for multi-channel (more than 2 channels) applications. Moreover these mostly involve simulation studies or real-life separations in controlled settings. In this paper some practical problems of multi-channel applications will be analyzed. A methodology is proposed to bypass the full fledged higher dimensional BSS problem by exploiting the relative spatial arrangement of microphones to achieve improved speech enhancement by selective 2-channel BSS. The concept is illustrated on a 3 source × 3 microphone setup in speech recognition experiments involving an acoustic scene with a human speaker and interfering noise sources.

1. Introduction

BSS routines for speech applications are designed to blindly invert an acoustic mixing matrix in multi-source/multi-microphone configurations [1, 2, 3]. However separations on real recorded data are hard to achieve whenever the relative source-microphone arrangements lead to singular or otherwise ambiguous source mixing situations. Even if an exact singularity is not encountered, situations in proximity of this singularity will lead to bad separation performance.

The dependence of BSS separation performance on the spatial allocation of sources and sensors has been well documented [2, 5, 7]. For most frequency domain BSS algorithms, this is especially important since solution of the permutation problem depends on accurate source DOA estimation. For low dimensional source/microphone setups, the limitations of source/microphone arrangements are relatively easy to investigate and guidelines can be devised to work around them. However for higher dimensional setups, the situation is less transparent because of the multitude of possible relative source/mic and mic/mic configurations. For some controlled settings, multi-dimensional source separations have been reported under appropriate assumptions [4]. Given the lack of theoretical guidelines in this case, we propose in this paper a methodology building on 2-channel BSS units performed on selected microphone pairs in a higher dimensional microphone configuration.

2. Singularity

We will first illustrate scaling issues of the singularity problem. For the 2 source × 2 microphone case, recorded microphone signals \( y_i, i \in (1,2) \), can be modeled by

\[
\begin{bmatrix}
    y_1 \\
    y_2
\end{bmatrix} =
\begin{bmatrix}
    x_1 \\
    x_2
\end{bmatrix},
\]

where \( x_j, j \in (1,2) \), denote the source signals and the acoustic mixing matrix \( M \) is given by

\[
M = \begin{bmatrix}
    H_{11} & H_{12} \\
    H_{21} & H_{22}
\end{bmatrix}.
\]

The singularity condition \( \det(M) = 0 \) yields the relation

\[
H_{11} H_{22} - H_{12} H_{21} = 0
\]

Conditions in space approximately satisfying this relation are for example both sources located close to each other on one side of the mics (Figure 1, Case a) or both sources on the dividing line between the 2 mics (Figure 1, Case b). Case c) in Figure 1 is not a singular situation per se but represents a poorly conditioned scenario for the direction of arrival estimation [5]. Since many BSS approaches rely on direction of arrival estimation to solve the frequency domain permutation problem, this situation will result in similar poor separation than singular setups. A detailed study of singular setups is given in [7] for example.

So singularities can occur depending on the relative source-mic configuration and for the 2 × 2 case, they are fairly easy to identify. For the 3 source × 3 microphone case, the mixing matrix is given by

\[
M = \begin{bmatrix}
    H_{11} & H_{12} & H_{13} \\
    H_{21} & H_{22} & H_{23} \\
    H_{31} & H_{32} & H_{33}
\end{bmatrix}.
\]
leading to the condition

\[ H_{11} H_{22} H_{33} - H_{11} H_{32} H_{23} - H_{21} H_{12} H_{33} + H_{21} H_{13} H_{32} + H_{31} H_{12} H_{23} - H_{31} H_{13} H_{22} = 0 \]  

(5)

When the time-frequency mask is not designed appropriately and only works with sources exhibiting sparse time-frequency characteristics. Moreover these techniques heavily rely on knowledge of the number of sources involved.

Instead we propose a nonlinear microphone setup where a BSS algorithm is run on pairs of microphones followed by a conventional single channel speech enhancement approach. The method should benefit from an efficient microphone configuration to handle both sparse and non sparse time-frequency interfering point sources. Moreover this technique relies on little a priori information.

3. Nonlinear microphone setup

Under these circumstances it seems like working with a lower dimensional BSS approach applied to microphone pairs and robust additional post processing provides a more practical alternative to solving full dimensional BSS to all microphone recordings at once. An optimal mic pair setup should be chosen by using the knowledge of situations described by Figure 1 and avoiding them.

One way to approach the problem is to work with an undercomplete BSS algorithm using some probabilistic speech models to reconstruct underlying source signals [9, 10], or use binary masking techniques [11]. The former method requires a priori information about the source signals and suffers from computational load problems in the clustering step. The latter technique on the other hand suffers from inherent musical artifacts when the time-frequency mask is not designed appropriately and only works with sources exhibiting sparse time-frequency characteristics. Moreover these techniques heavily rely on knowledge of the number of sources involved.

Instead we propose a nonlinear microphone setup where a BSS algorithm is run on pairs of microphones followed by a conventional single channel speech enhancement approach. The method should benefit from an efficient microphone configuration to handle both sparse and non sparse time-frequency interfering point sources. Moreover this technique relies on little a priori information.

4. Experiments

To illustrate this reduced complexity approach, the acoustical scene depicted in Figure 2 was studied. A speech source to be enhanced was localized in front of microphone 2 and all sources were about 0.7-2 m from the closest microphone (Figure 2). The 3 microphones were located at a distance of 5-15 cm with respect to each other. A radio news channel was played from speaker position source 1 while a dense and colored but quasi-stationary noise was played from position source 3. A human speaker located at source position 2 was uttering words from a 20 word vocabulary list.

We considered a number of different BSS approaches which however showed similar results. An extension of the BSS approach based on the typical algorithm used in our study [3, 8] was finally employed. BSS separation was followed by a minimum statistics type post processing approach [12] on separated output channel 2 for the 2-channel BSS cases. In speech recognition experiments, we compared the 3-channel BSS result with the results for microphone pairs (1,2), (2,3) and (1,3) and the one obtained by spectral subtraction. We indicate speech recognition rates as the performance measure rather than SIR or distortion rates which give a difficult overall picture of the speech enhancement achieved.

The speech recognition results are shown in Table 1. 4 subjects uttered a total of 160 words from a list.
of 20 vocabulary words in acoustic scenarios where either one non-stationary source was played from source position 1 or both source 1 and source 3 were active. The speech recognizer was an off-the-shelf commercial system with specifiable vocabulary operating at 8 kHz sampling frequency. For the original/spectral subtraction study case, the signal recorded at microphone 2 was recognized with/without single channel processing respectively. Again a minimum statistics type approach was used for spectral subtraction [12]. In processing case BSS 12, the separated output at channel 2 was analyzed while the best recognition result for both channels output by BSS is shown for mic pair cases 23 and 13. For 3-chan BSS, the separated output at channel 2 is considered.

Figures 3, 4 and 5 illustrate the speech enhancement achieved with either method. One can guess that methods BSS 23, BSS 13 and 3-chan BSS achieve similar SNR, so the difference in speech recognition scores among the various cases is due to various degrees of reverberation and distortion artifacts associated with the different microphone recording choices. The SNR achieved by method BSS 12 is however clearly higher than the original and spectral subtraction case SNR.

<table>
<thead>
<tr>
<th>Case</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>60.0</td>
</tr>
<tr>
<td>Spec Sub</td>
<td>70.6</td>
</tr>
<tr>
<td>BSS 12</td>
<td>81.2</td>
</tr>
<tr>
<td>BSS 23</td>
<td>74.4</td>
</tr>
<tr>
<td>BSS 13</td>
<td>73.8</td>
</tr>
<tr>
<td>3-chan BSS</td>
<td>79.4</td>
</tr>
</tbody>
</table>

Table 1: Speech recognition accuracy for various processing methods:
Original=recorded microphone signal 2;
Spec Sub=spectral subtraction on microphone signal 2;
BSS 12= BSS applied to mic pair (1,2) followed by Single Channel Speech Enhancement(SCSE);
BSS 23= BSS applied to mic pair (2,3) + SCSE;
BSS 13= BSS applied to mic pair (1,3) + SCSE;
3-chan BSS= simultaneous 3 channel BSS

5. Discussion

It is observed that although the 3-channel BSS approach yields very good performance in this case, it has been achieved at the expense of significantly higher computational load (required convergence iterations, 6 times the number of filter coefficients of 2-channel approach). On the other hand, choosing 2-channel BSS applied to the right pair (1,2) achieves better results while choices (2,3) and (1,3) are not appropriate although still better than spectral subtraction. The microphone choice (1,3) is not suitable since source 1 and 3 lie close to the supporting axis of mic 1 and 3. Also none of these mics is close to the human speaker and the separated signals therefore contain more reverberation and distortion as a result of an initial worse SNR. The choice (2,3) was better than (1,2) since the highly nonstationary noise signal from source 1 was more efficiently removed using 2-channel BSS applied to mic pair (1,2).

Determining the best microphone pair for 2-channel BSS in a nonlinear microphone setup can thus yield better performance than a brute-force complete multichannel BSS approach while requiring considerably less computational load. However the appropriate microphone setup as well as the best mic pair choice in this configuration are critical elements for robust performance and will require some form of a priori knowledge about the acoustical scene.

6. Conclusions

A 3-channel BSS approach was compared to 2-channel BSS followed by single channel post processing on selected microphone pairs in speech recognition experiments in a real-life acoustic setting with one human speaker and two interfering point sources. The results indicate that the 2-channel BSS method applied to a suitable microphone pair yields better enhancement than the higher complexity 3-channel BSS approach. However the benefit of reduced computational burden and better understanding of singular situations in the selective 2-channel BSS method when compared to combined higher order microphone BSS requires additional information about the acoustical scene to determine the best microphone pair. This study suggests using a nonlinear microphone setup and an appropriate pair selection criterion will help address complex auditory scenes and can provide significant speech enhancement given reasonable computational resources.

7. References


Figure 3: Examples of noisy speech commands for original (upper) and after spectral subtraction (lower)

Figure 4: Examples of processed speech commands after BSS applied to microphone pair (1,2) (upper) and pair (2,3) (lower)

Figure 5: Examples of processed speech commands after BSS applied to microphone pair (1,3) (upper) and output at mic channel 2 of simultaneous 3 channel BSS (lower)