Multi-channel speech enhancement using sparse coding on local time-frequency structures

Zhiyuan Zhou[1], *Zhaogui Ding*[1], *Weifeng Li*[1],
Zhiyong Wu[2], *Longbiao Wang*[3], *and Qingmin Liao*[1]

[1]Department of Electronic Engineering / Graduate School at Shenzhen, Tsinghua University, China
[2]Department of Computer Science / Graduate School at Shenzhen, Tsinghua University, China
[3]Nagaoka University of Technology, Japan

10.21437/Interspeech.2014-580

Abstract

A novel multi-channel speech enhancement technique is proposed in the present paper. We focus on the local sparsities of speech signals in contrast to the conventional beamforming and blind source seperation methods. The technique utilizes the difference of local structures in temporary-frequency domain between the target speech and interfering signals for enhancing the target speech. We first estimate the local structures of the speech and noise signals at each time-frequency bin to form a local dictionary, and then recover the clean speech via sparse coding. The proposed algorithm is simple to implement and requires no prior knowledge of speech and noise. Our experimental evaluations demonstrate that the proposed method can suppress interferer and meantime preserve target speech more than some conventional methods.

Index Terms: microphone arrays, speech enhancement, sparse coding

1. Introduction

Speech quality and intelligibility often significantly degrade in the presence of background interfering noise. Consequently, modern communications systems employ speech enhancement algorithms at the preprocessing stage prior to further processing (e.g., coding, speech recognition). Speech enhancement algorithms have been attractive research in the past three decades. Multi-channel speech enhancement techniques take advantage of the availability of multiple signal inputs. Due to the possibility to perform spatial filtering, multi-channel speech enhancement algorithms have a better ability to increase speech quality as well as intelligibility of speech in noise [1].

The widely used multi-channel methods are based on microphone array beamforming techniques. The simplest technique is using the *delay-and-sum* beamformer, which synchronizes the target signal from a particular direction. Other more sophisticated beamforming methods, such as the superdirective beamformer [2] and Generalized Sidelobe Canceller (GSC) [3], optimize the beamformer to produce a spatial pattern with a dominant response for the location of interest. However, problems such as "signal leakage" [4]inherent in GSC have limited its effectiveness, and also a persistent problem with microphone arrays has been poor low-frequency directivity for practical array dimensions [5]. Other methods based on *blind source separation* (BSS) [6] or *independent component analysis* (ICA) [7] assume statistical independence between the target and noise signals which is not always true. Moreover, permutaion and scaling ambiguity problems limit their performance. Recently some attempt to combine beamforming and blind source separation techniques appeared in the literatures [8][9].

In this paper we introduce a novel multi-channel speech enhancement method based on enhancing the target speech via sparse coding [10]. Since clean speech and many kinds of interfering signals contain different structures, their structured components can be sparsely coded in coherent dictionaries separately. For instance, [11] directly learned the speech and noise dictionaries and [12] employed the exemplars of target speech and noise signals used for sparse coding. However rather than using speech and noise dictionaries separately and globally, we adopt a beamforming configuration to generate a local dictionary jointly consisting of local structures of both speech and interfering signals. Then the noisy signal can be sparsely represented over the generated local dictionary, and finally the local structure component of clean speech can be obtained by multiplying the local speech dictionary with its corresponding sparse coefficients. Our method novelly incorporates the knowledge of spatial difference of multiple microphone inputs and local temporary-frequency sparse coding, and does not assume any statistics of speech data. We demonstrate that our proposed technique provides effective nulling of the noise source, without the signal cancellation problems in conventional beamformings. Moreover, our technique does not suffer from the source permutation and scaling ambiguities encountered in conventional BSS algorithms.

2. Problem statement

In this paper, we consider the problem of extracting a speech source s, contaminated by an interfering noise source i. The signals are recorded by N sensors arranged in an microphone array. The target speech source and interfering source are not directly observable, but the positions of sources (both speech and interfering source) as well as the positions of sensors are known. In short-time Fourier transform (STFT) domain, the array signal model is defined as:

$$\mathbf{x}(k,f) = s(k,f)\mathbf{d}_s(f) + i(k,f)\mathbf{d}_i(f), \qquad (1)$$

where $\mathbf{x}(k,f) = [x_1(k,f), ..., x_N(k,f)]^T$, $s(k,f)$ and $i(k,f)$ are the complex-valued STFTs of the corresponding time signals. $\mathbf{d}_s(f)$ and $\mathbf{d}_i(f)$ represent the array steering vectors which depend on the actual microphone and source locations. $f = 1, ..., F$ is a frequency bin index, and $k = 1, ..., K$ is a frame index. For a source located far from the array, it is common to make a plane wave assumption and the vectors can

This work was supported in part by Shenzhen Basic Research Grant JCYJ20120831165730913.

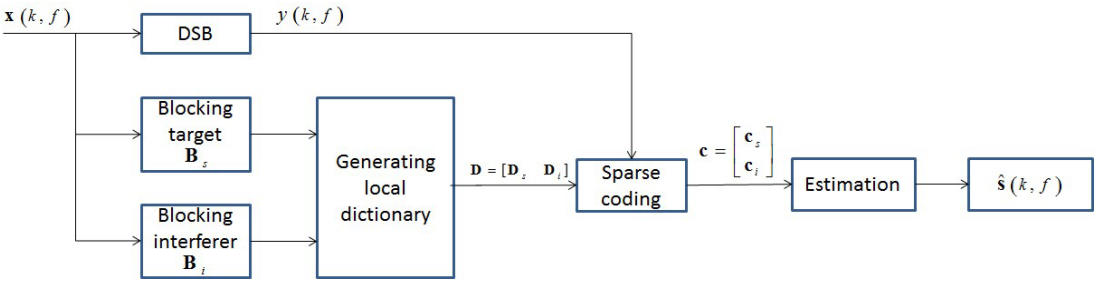

Figure 1: The block diagram of a multi-channel speech enhancement framework based on sparse coding

be calculated simply using the time delay of arrival (TDOA). $\mathbf{d}_s(f)$ and $\mathbf{d}_i(f)$ are given by:

$$\mathbf{d}_s(f) = [e^{-j2\pi f\tau_1/(LT_s)}, ..., e^{-j2\pi f\tau_N/(LT_s)}]^T, \quad (2)$$

$$\mathbf{d}_i(f) = [e^{-j2\pi f\gamma_1/(LT_s)}, ..., e^{-j2\pi f\gamma_N/(LT_s)}]^T, \quad (3)$$

where τ_n and γ_n denote the TDOAs of the target and the interfering signal respectively. L is the length of frame, and T_s is the sampling interval. Recovering the clean speech s from the noisy signal $\mathbf{x}$ without any prior knowledge about interfering signal is a difficult problem, for interfering signal can be stationary or nonstationary noise, even other speaker′s speech.

3. Proposed method

3.1. Local time-frequency structure of the signal

A key ingredient of this work is to represent the signal over an over-complete dictionary via sparse coding. If both the speech and interferer dictionaries are coherent to its respective structured component in the mixture signal, sparse coding is able to separate the mixture signal into its structured components using a learned dictionary [11]. The method proposed here is based on local time-frequency structure of the signal. We define the local frequency structure (LFS) $\mathbf{y}(k, f)$ of the signal y in time-frequency domain at bin index (k, f) as follows:

$$\mathbf{y}(k, f) = [|y(k, f - P)|, ..., |y(k, f)|, ..., |y(k, f + P)|]^T, \quad (4)$$

where $|y(k, f)|$ denotes the magnitude of $y(k, f)$. Then the local time-frequency structure (LTFS) $\mathbf{Y}(k, f)$ of the signal y at (k, f) is defined by:

$$\mathbf{Y}(k, f) = [\mathbf{y}(k - Q, f), ..., \mathbf{y}(k, f), ..., \mathbf{y}(k + Q, f)], \quad (5)$$

An illustration of LFS and LTFS of the signal at bin index (k, f) is showed in Fig. 2.

3.2. Proposed framework

The framework of our speech enhancement system is shown in Fig. 1. Firstly we transform the mixture time-domain signal into STFT domain and use delay-and-sum beamformer (DSB) for enhancing target signal. DSB is the simplest beamforming algorithm, and uses only the geometrical knowledge to enhance target signal. Each microphone output is weighted by frequency domain coefficients, and the beamformer output is the sum of each weighted microphone output:

$$y(k, f) = \frac{1}{N}\mathbf{d}_s^H(f)\mathbf{x}(k, f), \quad (6)$$

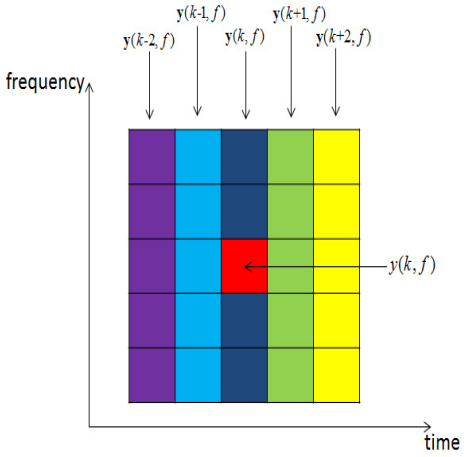

Figure 2: An illustration of LFS and LTFS of the signal at bin index (k, f), and LTFS can be expressed as $\mathbf{Y}(k, f) = [\mathbf{y}(k - 2, f), \mathbf{y}(k - 1, f), \mathbf{y}(k, f), \mathbf{y}(k + 1, f), \mathbf{y}(k + 2, f)]$, where $P = 2, Q = 2$.

where $\mathbf{d}_s(f)$ is the steering vector defined in Eq. (2).

The upper branch, DSB-enhanced signal $y(k, f)$, consists of both the target speech and interfering noise. In the lower branch, we design the following two dictionaries to approximate $y(k, f)$. More specifically, we generate a local dictionary $\mathbf{D}(k, f)$ that consists of speech signal dictionary $\mathbf{D}_s(k, f)$ and interferering dictionary $\mathbf{D}_i(k, f)$ using blocking matrix, as described in detail in Sec. 3.3. Then the LFS of firstly enhanced speech and the generated local dictionary are ready to be fed to following speech processing module to represent speech with sparse coding, as described in detail in Sec. 3.4. Through estimating, we would obtain the LFS of enhanced speech in frequency-domain. Finally we re-synthesize the final enhanced speech in time-domain using the phase of y.

3.3. Local dictionary generation

In this section, we describe the method to generate the local dictionary which consists of a target signal local dictionary and interferer local dictionary. Both should be more coherent to their own local structure. We use local structures of the speech signal and interfering signal as local dictionaries, i.e.,

$$\mathbf{D}_s(k, f) = \mathbf{S}(k, f), \ \mathbf{D}_i(k, f) = \mathbf{I}(k, f). \quad (7)$$

As speech and interfering sources are not directly observable, we need to estimate their local structures at time-frequency

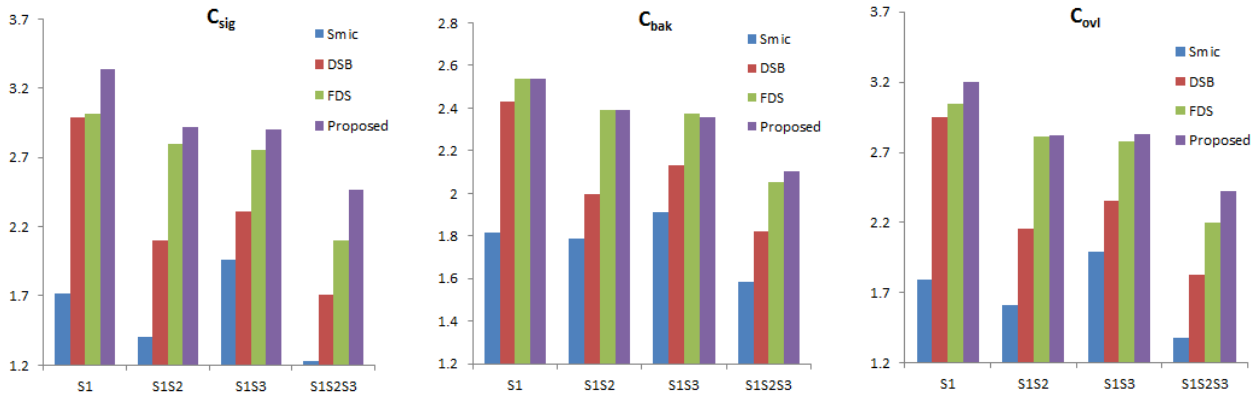

Figure 3: C_{sig}, C_{bak} and C_{ovl} scores obtained in various scenarios.

bin index (k, f). Firstly we use two blocking matrix $\mathbf{B}_s(f)$ and $\mathbf{B}_i(f)$ to block the target speech signal and interfering signal respectively. The blocked signals can be expressed as:

$$\mathbf{z}_s(k,f) = \mathbf{B}_s^H(f)\mathbf{x}(k,f), \tag{8}$$

$$\mathbf{z}_i(k,f) = \mathbf{B}_i^H(f)\mathbf{x}(k,f), \tag{9}$$

where $\mathbf{z}_s(k,f) = [z_{s1}((k,f), ..., z_{s(N-1)}(k,f)]^T \in \mathbb{C}^{N-1}$ and $\mathbf{z}_i(k,f) = [z_{i1}(k,f), ..., z_{i(N-1)}(k,f)]^T \in \mathbb{C}^{N-1}$. $\mathbf{B}_s(f), \mathbf{B}_i(f) \in \mathbb{C}^{N\times(N-1)}$ are orthogonal to $\mathbf{d}_s(f)$ and $\mathbf{d}_i(f)$ respectively, such that

$$\mathbf{B}_s^H(f)\mathbf{d}_s(f) = \mathbf{0},\ \mathbf{B}_i^H(f)\mathbf{d}_i(f) = \mathbf{0}. \tag{10}$$

The blocking matrix can be calculated with an orthogonalization technique, such as singular value decomposition (SVD), QR factorization or Gram-Schmidt orthogonalization technique.

Since $\mathbf{z}_s(k,f)$ mainly contains interfering signal and $\mathbf{z}_i(k,f)$ mainly contains speech signal, we can adopt $\mathbf{z}_s(k,f)$ and $\mathbf{z}_i(k,f)$ as local time-frequency structures (LTFS) of the interfering and target signals respectively:

$$\mathbf{S}(k,f) \approx \mathbf{Z}_i(k,f),\ \mathbf{I}(k,f) \approx \mathbf{Z}_s(k,f), \tag{11}$$

where

$$\mathbf{Z}_s(k,f) = [\mathbf{Z}_{s1}(k,f)\ \ \mathbf{Z}_{s2}(k,f)\ ...\ \ \mathbf{Z}_{s(N-1)}(k,f)], \tag{12}$$

$$\mathbf{Z}_i(k,f) = [\mathbf{Z}_{i1}(k,f)\ \ \mathbf{Z}_{i2}(k,f)\ ...\ \ \mathbf{Z}_{i(N-1)}(k,f)]. \tag{13}$$

$\mathbf{Z}_{sn}(k,f)$ and $\mathbf{Z}_{in}(k,f)$ denote LTFS of the speech-blocked signal z_{sn} and interferer-blocked signal z_{in}, respectively. According to Eqs (7) and (11), we have

$$\mathbf{D}_s(k,f) \approx \mathbf{Z}_i(k,f),\ \mathbf{D}_i(k,f) \approx \mathbf{Z}_s(k,f). \tag{14}$$

The final dictionary is constructed by simply concatenating the speech and interferer local dictionaries, i.e.,

$$\mathbf{D}(k,f) = [\mathbf{D}_s(k,f)\ \ \mathbf{D}_i(k,f)]. \tag{15}$$

3.4. Speech enhancement with sparse coding

As the upper branch enhanced signal $y(k,f)$ consists of both speech and interfering signals, we approximate $y(k,f)$ using the generated local dictionary, which is composed of the estimated local structures of speech and interfering signals. We employ sparse coding to achieve this goal. The aim of sparse coding is to represent a signal as a linear combination of only a few signal prototypes. We code each local frequency structure (LFS) $\mathbf{y}(k,f)$ in a local dictionary $\mathbf{D}(k,f)$ using LASSO regression

$$\begin{aligned} &\arg\min_{\mathbf{c}(k,f)} \|\mathbf{y}(k,f) - \mathbf{D}(k,f)\mathbf{c}(k,f)\|_2 \\ &= \arg\min_{\mathbf{c}(k,f)} \|\mathbf{y}(k,f) - [\mathbf{D}_s(k,f)\ \ \mathbf{D}_i(k,f)] \begin{bmatrix}\mathbf{c}_s(k,f)\\ \mathbf{c}_i(k,f)\end{bmatrix} \|_2 \\ &\text{subject to } \|\mathbf{c}(k,f)\|_1 \le \lambda, \end{aligned} \tag{16}$$

where $\|\cdot\|_2$ and $\|\cdot\|_1$ denote l^2-norm and l^1-norm, respectively. λ is a sparsity parameter.

Then we can obtain the LFS of enhanced speech as

$$\widehat{\mathbf{s}}(k,f) = \mathbf{D}_s(k,f)\mathbf{c}_s(k,f). \tag{17}$$

We take the middle element of $\widehat{\mathbf{s}}(k,f)$ as the estimated speech magnitude $\hat{s}(k,f)$. The final waveform signal is reconstructed by overlapping and adding successive output frames.

3.5. Related to other works

Although both [11] and [12] employ sparse coding for speech enhancement and automatic speech recognition respectively, our method is based on multiple microphone array inputs. Moreover our dictionary generation is in the context of beamforming configurations, which are totally different from them. [13] proposed a multi-channel speech enhancement algorithm based on convex optimization and pause detection of the speech sources. Our method is straightly based on beamforming configuration and does not need pause detection.

4. Experiments

For the experiments, we use a subset of The Multichannel Overlapping Numbers Corpus (MONC) [15]. The subset data contain 100 utterances for each scenarios (S1, S1S2, S1S3, S1S2S3). S1 records audio files for the desired speaker only (no overlapping speech) scenario. S1S2 and S1S3 record audio files for the desired speaker and one competing speaker with different location. S1S2S3 records audio files for the desired speaker and two competing speakers scenario. We use the center microphone signal as SMic signal, which can be considered a mixture

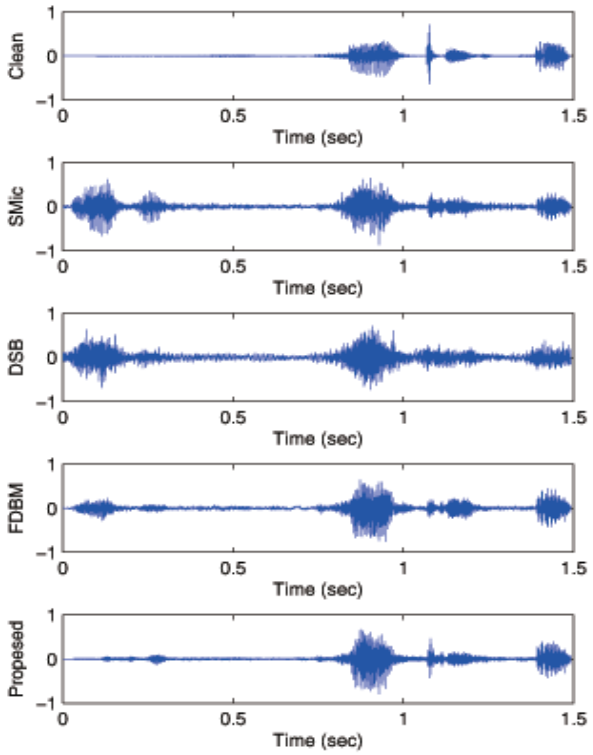

Figure 4: Example waveforms of clean speech, SMic speech and speech enhanced with DSB, FDBM, and our proposed methods (from top to down).

of the target and interfering speech. We set $P = 1$ and $Q = 3$ in our experiments.

To evaluate the effectiveness of the proposed method, we perform a series of speech enhancement experiments. We compare the performance of our proposed method with the delay-and-sum beamforming (DSB), the frequency-domain binary masking beamforming (FDBM) [14], as well the baseline using a single distance microphone (SMic). We choose DSB and FDBM because they are both proven to be effective for overlapping speech enhancement [14], and FDBM is one of the state-of-the-art methods.

Enhancement performance is measured by composite objective measures C_{sig}, C_{bak} and C_{ovl} [16], which are obtained by linearly combining existing objective measures to form new measures : (a) C_{sig} is formed for target signal distortion by linearly combining the WSS, PESQ, and LLR measures; (b) C_{bak} is formed for background noise distortion by linearly combining the WSS, PESQ, and segSNR measures; (c) C_{ovl} is formed for overall quality by linearly combining the WSS, PESQ, and LLR measures. These new composite measures show moderate improvements over the existing objective measures [16]. The ratings are based on the $1-5$ MOS scale, range from 1 (bad) to 5 (excellent). Fig. 3 shows C_{sig}, C_{bak} and C_{ovl} scores for the various scenarios. Clearly, our method outperforms the DSB and FDBM in all scenarios. The proposed method yields an average improvement of above 1 compared to the scores obtained using a single distance microphone signal (SMic).

Speech waveforms are a useful tool for evaluating speech enhancement algorithms. Example waveforms of clean and noisy speech and also those of the outputs of the DSB, FDBM, and proposed method are presented in Fig. 4. The figure shows that the interferers was suppressed to a greater degree with the proposed method than with others. Meantime the proposed method can preserve the target speech more than other methods.

5. Conclusions

In this paper, we proposed a new method to enhance speech using sparse coding. We focus on local structure of the signal in time-frequency domain. Experiments on the multichannel numbers corpus (MONC) [15] have illustrated that the proposed method can recover the target speech signal more accurately than the conventional beamforming methods.

6. References

[1] Dines, J., Bourlard, H. and Li, W.,"MLP-based Log Spectral Energy Mapping for Robust Overlapping Speech Recognition[R]", IDIAP, 2007.

[2] Brandstein, M. and Ward, D. (Eds.),"Microphone arrays: signal processing techniques and applications", Springer, 2001.

[3] Griffiths, L. J. and Jim, C. W.,"An alternative approach to linearly constrained adaptive beamforming", Antennas and Propagation, IEEE Transactions on, 30(1), 27-34, 1982.

[4] Herbordt, W., Buchner, H. and Kellermann, W.,"An acoustic human-machine front-end for multimedia applications", EURASIP Journal on Applied Signal Processing, 21-31, 2003.

[5] McCowan, I. A., Moore, D. C. and Sridharan, S.,"Near-field adaptive beamformer for robust speech recognition", Digital Signal Processing, 12(1), 87-106, 2002.

[6] Haykin, S.,"Unsupervised adaptive filtering", New York : Wiley, 2000.

[7] Lee, T. W.,"Independent component analysis", 27-66, Springer US, 1998.

[8] Kumatani, K., Gehrig, T., Mayer, U., Stoimenov, E., McDonough, J. and Wolfel, M.,"Adaptive beamforming with a minimum mutual information criterion", Audio, Speech, and Language Processing, IEEE Transactions on, 15(8), 2527-2541, 2007.

[9] Wang, L., Ding, H. and Yin, F.,"Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals", EURASIP Journal on Audio, Speech, and Music Processing, 2010(797962), 1-13, 2010.

[10] Cands, E. J., and Wakin, M. B.,"An introduction to compressive sampling", Signal Processing Magazine, IEEE, 25(2), 21-30, 2008.

[11] Sigg, C. D., Dikk, T. and Buhmann, J. M.,"Speech enhancement using generative dictionary learning", Audio, Speech, and Language Processing, IEEE Transactions on, 20(6), 1698-1712, 2012.

[12] Gemmeke, J. F., Virtanen, T. and Hurmalainen, A.,"Exemplar-based sparse representations for noise robust automatic speech recognition", Audio, Speech, and Language Processing, IEEE Transactions on, 19(7), 2067-2080, 2011.

[13] Yu, M., Ma, W., Xin, J. and Osher, S.,"Multi-Channel Regularized Convex Speech Enhancement Model and Fast Computation by the Split Bregman Method", Audio, Speech, and Language Processing, IEEE Transactions on, 20(2), 661-675, 2012.

[14] Maganti, H. K., Gatica-Perez, D. and McCowan, I.,"Speech enhancement and recognition in meetings with an audioCvisual sensor array", Audio, Speech, and Language Processing, IEEE Transactions on, 15(8), 2257-2269, 2007.

[15] Mccowan, I. A.,"The Multichannel Overlapping Numbers Corpus", Idiap resources available online: http://www. cslu. ogi. edu/corpora/monc. pdf, 2003.

[16] Hu, Y. and Loizou, P. C.,"Evaluation of objective quality measures for speech enhancement", Audio, Speech, and Language Processing, IEEE Transactions on, 16(1), 229-238, 2008.