2020 RISP International Workshop on Nonlinear Circuits,
Communications and Signal Processing (NCSP 2020)
Honolulu, Hawaii, USA, February 28 - March 2, 2020

Blind Source Separation with Low Latency for In-Car Communication

Tetsuya Ueda, Shota Inoue, Shoji Makino, Mitsuo Matsumoto, Takeshi Yamada

University of Tsukuba
1–1–1 Tennodai, Tsukuba, Ibaraki, Japan
E-mail: {t.ueda@mmlab.cs, s.inoue@mmlab.cs, maki@tara, makilab-research@tara, takeshi@cs}.tsukuba.ac.jp

Abstract

We examined blind source separation (BSS) with low latency in cars. In frequency-domain BSS, a buffer delay associated with a short-time Fourier transform (STFT) is inevitable. Thus, we need to shorten the STFT frame length. However, shortening of the STFT frame degrades the source separation performance because an assumption of instantaneous mixture model does not hold. On a similar research, the method of formulating a system for separating highly reverberant mixture signals using a frequency-domain convolutive mixture model has recently proposed. We paid attention to the method and considered an application of reducing the delay dependent on the STFT frame. We evaluated the effectiveness of the method in cars under a short reverberant condition. Experimental results confirmed that the method not only maintained its performance but also reduced the delay dependent on the STFT frame.

1. Introduction

Information and communication technology (ICT) is developing rapidly. Recently, several studies of the application of ICT in in-car communication (ICC) have been carried out [1, 2]. Owing to the background noise and seat arrangement, communication within a vehicle is difficult. In particular, backseat passengers feel uncomfortable listening to the sound from the speaker located near the driver seat. The situation can be improved by enhancing particular audio signals and transmitting them without delay. One of the common approaches to enhancing particular audio signals is blind source separation (BSS). BSS is a technique for separating individual source signals from recorded microphone array inputs without any prior information about source signals.

The most commonly used approach for overdetermined BSS (where the number of microphones is larger than that of sources) is independent component analysis (ICA) [3], which achieves source separation by assuming the statistical independence between the sources. Recently, a number of methods based on frequency-domain ICA have been developed [4, 5, 6], which provide flexibility in utilizing various models for the time–frequency representations of source signals and array responses. For example, independent low-rank matrix analysis (ILRMA) [6] adopts the non-negative matrix factorization (NMF) concept [7, 8] for source spectrogram modeling, which approximates each source power spectrogram as a linear combination of a limited set of spectral templates scaled by magnitudes varying with time.

In methods based on frequency-domain, there is a delay dependent on the STFT frame. This delay is the time for waiting for new signals so that the STFT frame buffer is filled. A high processing CPU and an efficient algorithm can not reduce the delay dependent on the STFT frame. Thus, in low latency methods based on frequency-domain, we need to shorten the STFT frame length. However, frequency-domain instantaneous mixture model such as ILRMA has an assumption that source separation performance degrades when the reverberation time is longer than the STFT frame length. Generally, the reverberation time in cars is short, but we can not ignore the reverberation problem if we shorten the STFT frame length. Although we can solve this problem by increasing the STFT frame length, it is difficult to separate sources with low latency owing to the delay dependent on the STFT frame.

On a similar research, the idea of formulating a system for separating highly reverberant mixture signals using a frequency-domain convolutive mixture model has recently been adopted in ILRMA [9]. We refer to this method as convolutive ILRMA, which has been shown to be effective for separating sound signals when the reverberation time is longer than the STFT frame length. In this paper, we applied convolutive ILRMA to low latency BSS in cars, where the reverberation time is shorter than the STFT frame length.

2. Frequency-domain convolutive ILRMA

2.1 Formulation based on frequency-domain instantaneous mixture model

We consider a determined situation where J source signals are observed by I microphones ($J = I$). Let $x_i(f, n)$ and $s_j(f, n)$ denote the STFT coefficients of the signal observed at the ith microphone and the jth source signal, where f and n are the frequency and time indices, respectively.

With the frequency-domain instantaneous mixture model, the relationship between the observed signals $\boldsymbol{x}(f,n) = [x_1(f,n),\dots,x_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I$ and source signals $\boldsymbol{s}(f,n) = [s_1(f,n),\dots,s_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I$ is written as

$$\boldsymbol{s}(f,n) = \boldsymbol{W}^{\boldsymbol{H}}(f)\boldsymbol{x}(f,n), \tag{1}$$

$$\boldsymbol{W}(f) = [\boldsymbol{w}_1(f),\dots,\boldsymbol{w}_I(f)] \in \mathbb{C}^{I\times I}, \tag{2}$$

where $\boldsymbol{W}^{\boldsymbol{H}}(f)$ is the separation matrix and $(\cdot)^{\boldsymbol{H}}$ denotes the Hermitian transpose.

Let us assume that $s_j(f,n)$ independently follows a zero-mean complex Gaussian distribution with variance $v_j(f,n) = \mathbb{E}[|s_j(f,n)|^2]$,

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n)|0, v_j(f,n)). \tag{3}$$

We also assume $s_j(f,n)$ to be independent from one source to the others. $\boldsymbol{s}(f,n)$ thus follows

$$\boldsymbol{s}(f,n) \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}(f,n)|0, \boldsymbol{V}(f,n)), \tag{4}$$

where $\boldsymbol{V}(f,n)$ is a diagonal matrix with diagonal entries $v_1(f,n),\dots,v_I(f,n)$. We further assume $v_j(f,n)$ as

$$v_j(f,n) = \sum_{k=1}^{K} h_{j,k}(f)u_{j,k}(n), \tag{5}$$

where $h_{j,k}(f) \geqq 0$ is the (j,k) element of the basis matrix and $u_{j,k}(n) \geqq 0$ is the (j,k) element of the activation matrix for the jth source. $k = 1,\dots,K$ are the numbers of the basis. In these assumptions, the negative log-likelihood of the parameters $\mathcal{V} = \{\boldsymbol{H},\boldsymbol{U}\}$ with $\boldsymbol{H} = \{h_{j,k}(f)\}_{j,k,f}$ and $\boldsymbol{U} = \{u_{j,k}(n)\}_{j,k,n}$, and $\mathcal{W} = \{\boldsymbol{W}^{\boldsymbol{H}}(f)\}_f$ given the observed mixture signals $\mathcal{X} = \{x_i(f,n)\}_{i,f,n}$ is given as

$$\mathcal{I}(\mathcal{W},\mathcal{V}|\mathcal{X}) \overset{c}{=} -2N\log\left|\det \boldsymbol{W}^{\boldsymbol{H}}(f)\right| + \sum_{f,n,j}\left(\log v_j(f,n) + \frac{|\boldsymbol{w}_j^{\boldsymbol{H}}(f)\boldsymbol{y}(f,n)|^2}{v_j(f,n)}\right), \tag{6}$$

where $\overset{c}{=}$ denotes equality up to constant terms.

This instantaneous formulation can not shorten the STFT frame length because source separation performance degrades when the reverberation time is longer than the STFT frame length.

2.2 Formulation based on frequency-domain convolutive mixture model

The idea of formulating a system for separating highly reverberant mixture signals using a frequency-domain convolutive mixture model has been proposed [9, 10] and shown to be effective for separating under the condition where the STFT frame length is not longer than the reverberation time. The relationship between the observed signals $\boldsymbol{x}(f,n)$ and sources $\boldsymbol{s}(f,n)$ is written as

$$\boldsymbol{s}(f,n) = \sum_{n'=0}^{N'} \boldsymbol{W}^{\boldsymbol{H}}(f,n')\boldsymbol{x}(f,n-n'), \tag{7}$$

where $\boldsymbol{W}(f,n')$, $0 \le n' \le N'$ are the coefficient matrices of size $I \times I$. N' is the length of the filter $\{\boldsymbol{W}^{\boldsymbol{H}}(f,n')\}_{f,n'}$. $\boldsymbol{W}^{\boldsymbol{H}}(f,0)$ is equivalent to a separation matrix of the instantaneous mixture model. When $\boldsymbol{W}^{\boldsymbol{H}}(f,0)$ is invertible, the dereverberated mixture signal $\boldsymbol{y}(f,n) = [y_i(f,n),\dots,y_I(f,n)]^{\mathsf{T}} \in \mathbb{C}^I$ and the source signal $\boldsymbol{s}(f,n)$ can be written as

$$\boldsymbol{y}(f,n) = \boldsymbol{x}(f,n) - \sum_{n'=1}^{N'} \boldsymbol{D}^{\boldsymbol{H}}(f,n')\boldsymbol{x}(f,n-n'), \tag{8}$$

$$\boldsymbol{s}(f,n) = \boldsymbol{W}^{\boldsymbol{H}}(f,0)\boldsymbol{y}(f,n), \tag{9}$$

where $\boldsymbol{D}^{\boldsymbol{H}}(f,n') = -(\boldsymbol{W}^{\boldsymbol{H}}(f,0))^{-1}\boldsymbol{W}^{\boldsymbol{H}}(f,n')$, $1 \le n' \le N'$. Equation (8) can be seen as a dereverberation process of the observed mixture signal $\boldsymbol{x}(f,n)$. $\mathcal{D} = \{\boldsymbol{D}^{\boldsymbol{H}}(f,n')\}_{f,n'}$ represents a dereverberation filter whose length is N'. Therefore, the negative log-likelihood of interest is a function of the dereverberation filter $\mathcal{D}$, separation matrices $\mathcal{W}$, and spectral parameters $\mathcal{V}$:

$$\mathcal{I}(\mathcal{D},\mathcal{W},\mathcal{V}|\mathcal{X}) \overset{c}{=} -2N\log\left|\det \boldsymbol{W}^{\boldsymbol{H}}(f)\right| + \sum_{f,n,j}\left(\log v_j(f,n) + \frac{|\boldsymbol{w}_j^{\boldsymbol{H}}(f)\boldsymbol{y}(f,n)|^2}{v_j(f,n)}\right) \tag{10}$$

2.3 Optimization process

We describe the optimization algorithm in this subsection. The objective function (10) is iteratively decreased using a coordinate descent method in which each iteration comprises the following three minimization steps:

$$\hat{\mathcal{V}} \leftarrow \underset{\mathcal{V}}{\operatorname{argmin}}\, \mathcal{I}(\mathcal{D},\mathcal{W},\mathcal{V}|\mathcal{X}), \tag{11}$$

$$\hat{\mathcal{W}} \leftarrow \underset{\mathcal{W}}{\operatorname{argmin}}\, \mathcal{I}(\mathcal{D},\mathcal{W},\mathcal{V}|\mathcal{X}), \tag{12}$$

$$\hat{\mathcal{D}} \leftarrow \underset{\mathcal{D}}{\operatorname{argmin}}\, \mathcal{I}(\mathcal{D},\mathcal{W},\mathcal{V}|\mathcal{X}). \tag{13}$$

The update rules of minimizing $\mathcal{V}$ are written as [8] ,

$$h_{j,k}(f) = h_{j,k}(f)\sqrt{\frac{\sum_n |s_j(f,n)|^2 u_{j,k}(n)v_j^{-2}(f,n)}{\sum_n u_{j,k}(n)v_j^{-1}(f,n)}}, \tag{14}$$

$$u_{j,k}(n) = u_{j,k}(n)\sqrt{\frac{\sum_f |s_j(f,n)|^2 h_{j,k}(f)v_j^{-2}(f,n)}{\sum_f h_{j,k}(f)v_j^{-1}(f,n)}}. \tag{15}$$

The update rules of $\mathcal{W}$ can easily be derived on the basis of the IP method [4],

$$\boldsymbol{w}_j(f) \leftarrow (\mathbf{W}^{\boldsymbol{H}}(f)\boldsymbol{\Sigma}_j(f))^{-1}\boldsymbol{e}_j, \tag{16}$$

$$\boldsymbol{w}_j(f) \leftarrow \frac{\boldsymbol{w}_j(f)}{\sqrt{\boldsymbol{w}_j^{\boldsymbol{H}}(f)\boldsymbol{\Sigma}_j(f)\boldsymbol{w}_j(f)}}, \tag{17}$$

where $\boldsymbol{\Sigma}_{y/v_j}(f) = (1/N)\sum_n \boldsymbol{y}(f,n)\boldsymbol{y}^{\boldsymbol{H}}(f,n)/v_j(f,n)$ and $\boldsymbol{e}_j$ denotes the jth column of the $I \times I$ identity matrix.

To obtain the update rules for $\mathcal{D}$, we vectorize $\boldsymbol{D}(f,n')$ as

$$\begin{aligned}\boldsymbol{d}(f) &= \text{vec}(\{\boldsymbol{D}(f,n')\})\\ &= [\boldsymbol{d}_1^{\mathsf{T}}(f,1),\dots,\boldsymbol{d}_I^{\mathsf{T}}(f,1),\boldsymbol{d}_1^{\mathsf{T}}(f,2),\dots,\boldsymbol{d}_I^{\mathsf{T}}(f,2),\\ &\quad \dots,\boldsymbol{d}_1^{\mathsf{T}}(f,N'),\dots,\boldsymbol{d}_I^{\mathsf{T}}(f,N')]^{\mathsf{T}} \in \mathbb{C}^{I^2N'},\end{aligned} \tag{18}$$

where $\boldsymbol{d}_i(f,n')$ is the ith column of $\boldsymbol{D}(f,n')$. The update rules for $\boldsymbol{d}^*(f)$ are given as

$$\begin{aligned}\boldsymbol{d}^*(f) \leftarrow &\left(\sum_n \boldsymbol{X}^{\boldsymbol{H}}(f,n)\boldsymbol{\Sigma}_{w/v(f,n)}\boldsymbol{X}(f,n)\right)^{-1}\\ &\times\left(\sum_n \boldsymbol{X}^{\boldsymbol{H}}(f,n)\boldsymbol{\Sigma}_{w/v(f,n)}\boldsymbol{x}(f,n)\right),\end{aligned} \tag{19}$$

where $\boldsymbol{\Sigma}_{w/v(f,n)} = \sum_j \frac{\boldsymbol{w}_j(f)\boldsymbol{w}_j^{\boldsymbol{H}}(f)}{v_j(f,n)}$ and $(\cdot)^*$ represents the complex conjugate of $(\cdot)$. We write $\boldsymbol{X}(f,n)$ as

$$\begin{aligned}\boldsymbol{X}(f,n) = [&\boldsymbol{I}\otimes\boldsymbol{x}^{\mathsf{T}}(f,n-1),\boldsymbol{I}\otimes\boldsymbol{x}^{\mathsf{T}}(f,n-2),\dots,\\ &\boldsymbol{I}\otimes\boldsymbol{x}^{\mathsf{T}}(f,n-N')] \in \mathbb{C}^{I\times I^2N'}.\end{aligned} \tag{20}$$

Here, $\boldsymbol{I}$ stands for the identity matrix of size $I \times I$ and $\otimes$ stands for the Kronecker product.

Therefore, the proposed algorithm is summarized as follows:

1. Initialize $\mathcal{V}$, $\mathcal{W}$, and $\mathcal{D}$.
2. Repeat the following updates for each j, f, n until convergence.
   (a) Update $h_{j,k}(f)$ and $u_{j,k}(n)$ using (14), (15).
   (b) Update $\boldsymbol{w}_j(f)$ using (16), (17).
   (c) Update $\boldsymbol{d}^*(f)$ using (19).

2.4 Application of convolutive ILRMA to ICC

ICC systems must fulfill strict delay constraints to ensure that the amplified speech is not perceived as a distinguishable echo. To realize a comfortable situation, mouth-to-ear delays should be maintained below 12 ms [2]. Thus, it is necessary for practical applications to shorten the STFT frame length. However, shortening of the STFT frame degrades the source separation performance because the assumption of instantaneous mixture model does not hold.

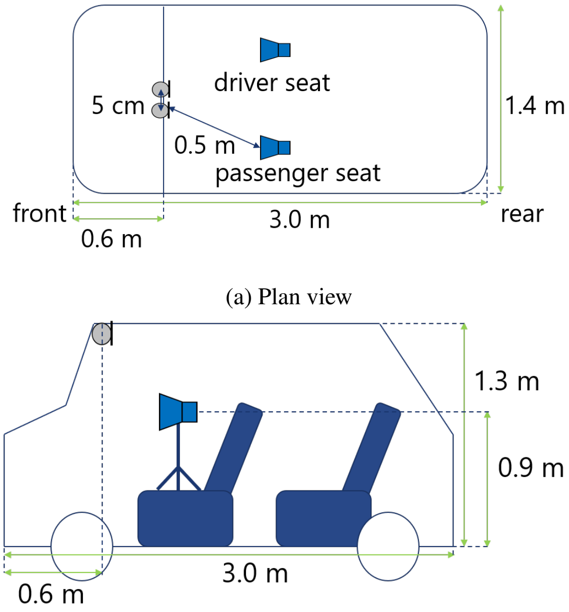
On the similar research in [9], convolutive ILRMA has been shown to be effective for separating sound signals under the condition where the reverberation time is longer than the STFT frame length. We paid attention to this advantage and considered an application of reducing the delay dependent on the STFT frame. In this paper, we applied convolutive ILRMA to low latency BSS in cars and investigated the performance in relation to the delay dependent on the STFT frame and the dereverberation filter length.

3. Experiments

To evaluate the effectiveness of convolutive ILRMA applied in cars, we conducted a source separation experiment. To investigate the improvement in the performance under the condition where the reverberation time is longer than the STFT frame length, we compared the performance in relation to the STFT frame length L and the dereverberation filter length N'. We used speech signals convolved with impulse responses recorded in a car. We took the average of the signal-to-distortion ratios (SDR) as the evaluation criteria [11].

3.1 Experimental conditions

In this experiment, we used clean data from a total of 10 speakers (6 male speakers and 4 female) and all 503 phoneme-balanced sentences contained in set B of the ATR digital speech database. By selecting two different speakers from the data set randomly, we obtained 10 patterns of source signals. We generated observed signals from the source signals by convoluting the impulse response recorded in a car. We resampled the source signals and the observed signals at 8 kHz. We measured the impulse responses using a time-stretched pulse in a car. The recording environment is shown in Fig. 1. We set one speaker at the driver seat, another speaker at a passenger seat, and microphones at a map lamp in the car. The reverberation time T_{60} in the car was 58 ms. In the experiment, we set L to $\{2, 4, 8, 16, 32\}$ ms and evaluated each SDR. We set N' in the range $0 \leq N' \leq 10$ because of a trade-off between N' and computational time of updating (19). $N' = 0$ is equivalent to separation in the instantaneous method. Moreover, we evaluated SDR at $L = 128$ ms and $N' = 0$ as baseline performance that satisfy the assumption of instantaneous mixture model. We set the STFT shift length to one-quarter of the STFT frame length. In [6], it was confirmed that ILRMA cannot achieve a good performance in the separation of speech signals when the number of bases is large. Thus, we set the number of bases to 1. We ran convolutive ILRMA for 50 iterations.



(a) Plan view

(b) Side view

Figure 1: Sound source and microphone layout in experiment

3.2 Results

Figure 2 shows the separation performance at each L. As the results show, convolutive ILRMA improved SDR at any L by choosing appropriate N'. Moreover, the results at $L \geq 4$ ms outperformed the baseline method at L = 128 ms and N' = 0. Therefore, we confirmed that convolutive ILRMA not only maintained its performance but also reduced the delay dependent on the STFT frame.

4. Conclusion

In this paper, we evaluated convolutive ILRMA (updating the separation matrices, the spectral parameters and the dereverberation filter iteratively) in cars. We conducted experiments in cars. The result of the experiments confirmed that the dereverberation filter is effective under a short reverberant condition. Therefore, convolutive ILRMA not only maintained its performance but also reduced the delay dependent on the STFT frame in cars.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 19H04131, SECOM Science and Technology Foundation, and Strategic Core Technology Advancement Program (Supporting Industry Program).

References

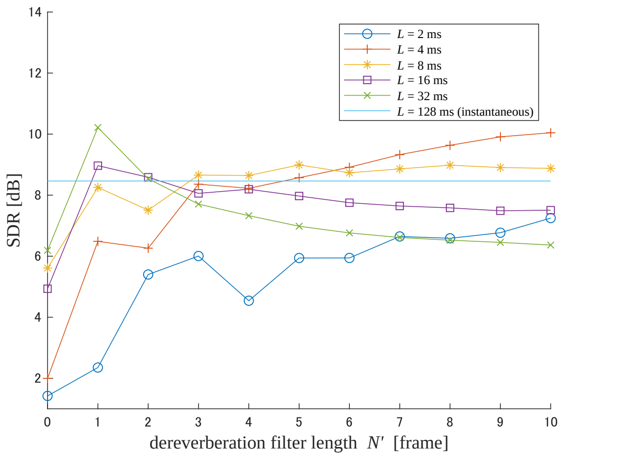


Figure 2: average SDR according to STFT frame length L and dereverberation filter length N'

[1] R. Landgraf *et al.*, “Can you hear me now? reducing the lombard effect in a driving car using an in-car communication system,” in Proc. Speech Prosody, pp. 479–483, 2016.

[2] A. Theiss *et al.*, “Instrumental evaluation of in-car communication systems,” in Proc. ITG, pp. 1–4, 2014.

[3] A. Hyvärinen *et al.*, “Independent component analysis: algorithms and applications”, Neural networks, Elsevier, vol. 13, no. 4, pp. 411–430, 2000.

[4] N. Ono, “Stable and fast update rules for independent vector analysis based on auxiliary function technique”, in Proc. WASPAA, pp. 189–192, 2011.

[5] H. Kameoka *et al.*, “Statistical model of speech signals based on composite autoregressive system with application to blind source separation”, in Proc. LVA/ICA, pp. 245–253, 2010.

[6] D. Kitamura *et al.*, “Determined blind source separation with independent low-rank matrix analysis”, *Audio Source Separation*, Springer, pp. 125–155, 2018.

[7] D. D. Lee *et al.*, “Algorithms for non-negative matrix factorization”, in Proc. NIPS, pp. 556–562, 2001.

[8] M. Nakano *et al.*, “Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence”, in Proc. MLSP, pp. 283–288, 2010.

[9] H. Kagami *et al.*, “Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization”, in Proc. ICASSP, pp. 31–35, 2018.

[10] T. Yoshioka *et al.*, “Blind separation and dereverberation of speech mixtures by joint optimization”, in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 1, pp. 69–84, 2010.

[11] E. Vincent *et al.*, “Performance measurement in blind audio source separation,” IEEE Trans. ASLP, vol. 14, no. 4, pp. 1462–1469, 2006.