TOWARD OVERCOMING FUNDAMENTAL LIMITATION IN FREQUENCY-DOMAIN BLIND SOURCE SEPARATION FOR REVERBERANT SPEECH MIXTURES

Lae-Hoon Kim and Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL, USA

ABSTRACT

Blind source separation can be implemented in the frequency domain using one-tap multiplication operation in each frequency bin, but only when the frame length is long enough to disregard temporal aliasing effects. If we take a short-time frequency transformation with a window shorter than a room reverberation time, the justification above does not hold anymore. In this paper, we present an appropriate representation in the short-time frequency domain. The suitability is justified by showing the equivalence with the original time domain approach under the overlap-add context. Experimental validation using a corpus synthesized by convolution with measured sets of room impulse responses is also provided.

1. INTRODUCTION

Frequency domain blind source separation (BSS) for reverberant speech mixtures has been extensively studied, because it can learn each frequency individually and selectively with much less computation under the assumption that convolution in the time domain can be represented as multiplication in the frequency domain. However, when we perform a frequency domain decomposition via short-time Fourier transform (STFT), we are often aware that the source separation filter should be longer enough than the conventional frame length (10 ms to 30 ms) for speech processing, because the reverberation time, typically 200-300 ms even in a small office environment, far exceeds the frame length. On the other hand, if we increase the frame length to make the filter length long enough under the same assumption, it results in decreasing the super-Guassianity of each frequency channel and consequently deteriorate the blind source separation performance. This fundamental limitation on frequency domain BSS has been reported [1], yet still recognized as unavoidable limitation.

In this paper, we propose a solution to overcome the limitation of the frequency domain BSS, which in fact starts from the important questions revoked on validity of the assumption on multiplication operation in STFT domain and presents the feed-forward representation proven to be a proper way with the shorter frame length than the reverberation time or more precisely the length of the inverse filter for the room impulse responses. Firstly, we develop an exact deconvolution operation in the STFT domain, demonstrating that it is similar to the time domain deconvolution in each frequency bin. Secondly, we extend the deconvolution scenario to the separation of reverberant mixtures by showing that a feed-forward network in the STFT domain is a proper implementation.

Remaining parts will be organized as follows. In section 2, we prove two main propositions regarding appropriate implementation of deconvolution and source separation in the STFT domain. In section 3, we experimentally validate the propositions we show in the section 2. In section 4, the impact of using the proper modeling in the domain of source separation will be discussed and summarized.

2. MAIN RESULTS

In this section, we provide two main propositions and corresponding proofs, which are developed based on overlap-addition method [2].

Proposition 1. *Deconvolution in the time domain by a filter with longer length than a frame length used for subband decomposition is equivalent to the deconvolution in each subband again.*

$$S(e^{j\omega_k}) = \sum_{l=0}^{N_W-1} W^l(e^{j\omega_k}) Y^{\cdot-l}(e^{j\omega_k}), \qquad (1)$$

where k is a frequency index, superscript $\cdot - l$ stands for the past frame, which is l frames before the current frame, and N_W represents the total number of frames to include the deconvolution filter length sufficiently, and the l^{th} tap subband domain convolution filter

$$W^l(e^{j\omega_k}) = \sum_{t=-\infty}^{\infty} w^l[t] e^{-j\omega_k t}, \qquad (2)$$

where

$$w^l[t] = w[t]win[lR - t], \quad -\infty < t < \infty, \qquad (3)$$

where win$[t]$ is a proper window function and from the overlap-add context

$$w[t] = \sum_{l=0}^{N_W-1} w^l[t]. \tag{4}$$

Proof. The equivalence between the subband domain deconvolution operation and the original time domain deconvolution can be demonstrated by taking an inverse Fourier transform on (1) and summing them over all possible frames, which is eventually shown to be equivalent to the results when we simply implement the deconvolution in the time domain with the original incoming signal and filter before applying subband decomposition. In the time domain, a deconvolution can be performed using a filter $w[t]$ with an incoming signal $y[t]$.

$$s[t] = \sum_{\tau=-\infty}^{\infty} y[\tau]w[t-\tau], \tag{5}$$

From the perspective of subband deconvolution, the deconvolved signal can be obtained as following.

$$\begin{aligned}
&\hat{s}[t] \\
&= \sum_{r=-\infty}^{\infty} \frac{1}{K} \sum_{k=0}^{K-1} \sum_{l=0}^{N_W-1} W^l(e^{j\omega_k}) Y^{r-l}(e^{j\omega_k}) e^{j\omega_k t} \\
&= \sum_{r=-\infty}^{\infty} \frac{1}{K} \sum_{k=0}^{K-1} \sum_{l=0}^{N_W-1} W^l(e^{j\omega_k}) \\
&\cdot \left[\sum_{\tau=-\infty}^{\infty} y[\tau] \mathrm{win}[(r-l)R-\tau] e^{-j\omega_k \tau} \right] e^{j\omega_k t} \\
&= \sum_{\tau=-\infty}^{\infty} y[\tau] \sum_{l=0}^{N_W-1} \left[\frac{1}{K} \sum_{k=0}^{K-1} W^l(e^{j\omega_k}) e^{j\omega_k(t-\tau)} \right] \\
&\cdot \sum_{r=-\infty}^{\infty} \mathrm{win}[(r-l)R-\tau] \\
&= \sum_{\tau=-\infty}^{\infty} y[\tau] \sum_{l=0}^{N_W-1} w^l[t-\tau] \sum_{r=-\infty}^{\infty} \mathrm{win}[(r-l)R-\tau] \\
&= \sum_{\tau=-\infty}^{\infty} y[\tau] w[t-\tau].
\end{aligned} \tag{6}$$

Note that with a careful choice of the window function, we can fulfill

$$\sum_{r=-\infty}^{\infty} \mathrm{win}[(r-l)R-\tau] = 1. \tag{7}$$

Therefore,

$$s[t] = \hat{s}[t], \tag{8}$$

and the subband domain deconvolution represented in (1) is a correct way of implementing a deconvolution in the subband domain. □

Proposition 2. *Multi-channel demixing of the convolutive mixture in the time domain with room impulse responses longer than a frame length used for subband decomposition is equivalent to the demixing of the convolutive mixture with delay of the frame shift in each subband.*

$$S_i(e^{j\omega_k}) = \sum_{j=1}^{N_S} \sum_{l=0}^{N_W-1} W_{i,j}^l(e^{j\omega_k}) Y_j^{\cdot -l}(e^{j\omega_k}), \tag{9}$$

where the subscript i means the i^{th} convolutive mixing output and j means the channel index. The number of the mixed sources N_S is not necessarily less than N_{mic}, but it is generally assumed so to perform source separation [3].

Proof. Proof of Proposition 2 is straightforward from Proposition 1. In the time domain, multi-channel deconvolutive demixing can be performed using a multi-channel filter $w_{i,j}[t]$ with an incoming signal $y_j[t]$.

$$s_i[t] = \sum_{j=1}^{N_S} \sum_{\tau=-\infty}^{\infty} y_j[\tau] w_{i,j}[t-\tau], \tag{10}$$

$$\begin{aligned}
&\hat{s}_i[t] \\
&= \sum_{r=-\infty}^{\infty} \frac{1}{K} \sum_{k=0}^{K-1} \sum_{j=1}^{N_S} \sum_{l=0}^{N_W-1} W_{i,j}^l(e^{j\omega_k}) Y_j^{r-l}(e^{j\omega_k}) e^{j\omega_k t} \\
&= \sum_{r=-\infty}^{\infty} \frac{1}{K} \sum_{k=0}^{K-1} \sum_{j=1}^{N_S} \sum_{l=0}^{N_W-1} W_{i,j}^l(e^{j\omega_k}) \\
&\cdot \left[\sum_{\tau=-\infty}^{\infty} y_j[\tau] \mathrm{win}[(r-l)R-\tau] e^{-j\omega_k \tau} \right] e^{j\omega_k t} \\
&= \sum_{j=1}^{N_S} \sum_{\tau=-\infty}^{\infty} y_j[\tau] \sum_{l=0}^{N_W-1} \left[\frac{1}{K} \sum_{k=0}^{K-1} W_{i,j}^l(e^{j\omega_k}) e^{j\omega_k(t-\tau)} \right] \\
&\cdot \sum_{r=-\infty}^{\infty} \mathrm{win}[(r-l)R-\tau] \\
&= \sum_{j=1}^{N_S} \sum_{\tau=-\infty}^{\infty} y_j[\tau] \sum_{l=0}^{N_W-1} w_{i,j}^l[t-\tau] \sum_{r=-\infty}^{\infty} \mathrm{win}[(r-l)R-\tau] \\
&= \sum_{j=1}^{N_S} \sum_{\tau=-\infty}^{\infty} y_j[\tau] w_{i,j}[t-\tau].
\end{aligned} \tag{11}$$

Therefore, with careful choice of the window function win,

$$s_i[t] = \hat{s}_i[t], \quad i = 1, 2, \cdots, N_{mic} \tag{12}$$

and (9) is a correct way of implementing the multi-channel deconvolutive demixing in the subband domain. □

3. EXPERIMENTAL VALIDATION

With one-tap multiplication only, frequency domain ICA is reported to converge to a spatial filter with a null in the direction of interference speech [4], i.e., to an optimal beamformer

with an additional null-forming constraint [5]. Recently, Kim et al. demonstrated improved reverberant speech separation, up to 29 dB C-weighted signal to interference ratio (SIR) [6], which is in fact about 20 dB more than the perforamnce of the conventional method with only one-tap multiplication, using a method that inspired Proposition 2 above. This section explores the implications of Proposition 2's formal guarantee, in particular, the ability to improve performance with more realistic room response assumptions.

3.1. Regularized feed-forward ICA (RFFICA) [6]

In [6], the feed-forward ICA has been introduced with the motivation of using not only the current frame but also the previous frames to maximize independence between separated speech stream with more similar super-Gaussianity of the original speech stream in each subband:

$$\vec{S} = \sum_{l=0}^{N_W-1} \mathbf{W}^l \vec{Y}^{\cdot-l}, \tag{13}$$

where $\vec{S} = [S_1\ S_2\ \cdots\ S_{N_S}]^T$, $\vec{Y}^{\cdot-l} = [Y_1^{\cdot-l}\, Y_2^{\cdot-l}\ \cdots\ Y_{N_{mic}}^{\cdot-l}]^T$ and

$$\mathbf{W}^l = \begin{bmatrix} W^l_{1,1} & W^l_{1,2} & \cdots & W^l_{1,N_S} \\ W^l_{2,1} & W^l_{2,2} & \cdots & W^l_{2,N_S} \\ \vdots & \vdots & \ddots & \vdots \\ W^l_{N_{mic},1} & W^l_{N_{mic},2} & \cdots & W^l_{N_{mic},N_S} \end{bmatrix}.$$

Note that (13) is a vector representation of (9).

The proposed update rule for the regularized feed-forward ICA (RFFICA) is given below:

$$\mathbf{W}^l = \mathbf{W}^l + \mu \left((1-\alpha) \cdot \Delta^l_{\text{ICA}} - \alpha \cdot \Delta^l_{\text{First stage}} \right), \tag{14}$$

where $l = 0, 1, \cdots, N_W - 1$, and N_W is the number of taps. The terms Δ^l_{ICA} and $\Delta^l_{\text{First stage}}$ represent the portion of the ICA update and the regularized portion on the first stage output.

$$\Delta^l_{\text{ICA}} = \mathbf{W}^l - \left\langle g\left(\vec{S}^{\cdot-(N_W-1)}\right) \vec{\hat{Y}}^{\cdot-l\,H} \right\rangle_t, \tag{15}$$

$$\vec{\hat{Y}} = \sum_{l=0}^{N_W-1} \mathbf{W}^{N_W-1-l\,H} \vec{S}^{\cdot-l}, \tag{16}$$

$$\Delta^l_{\text{First stage}} = \left\langle \vec{Y}^{\cdot-l}|_{\text{Ref}} \left(\vec{S}|_{\text{Ref}} - \vec{S}_{\text{First stage}} \right)^H \right\rangle_t, \tag{17}$$

where $< \cdot >_t$ represents time averaging, $(\cdot - l)$ represents l sample delay, $\vec{S}|_{\text{Ref}}$ is the first stage output vector for regularization, and $|_{\text{Ref}}$ represents the reference channels. The penalty term has been only applied to the channel where the references are assigned; the other entries for the mixing matrix are set to zero so that the penalty term vanishes on those channel updates. For initialization of the subsequent filters, we modeled the dereverberation process as exponential attenuation:

$$\mathbf{W}^l_{\text{ini}} = \exp(-\beta l) \cdot \mathbf{I}, \tag{18}$$

where $\mathbf{I}$ is an identity matrix, β is selected to model the average reverberation time, and l is the tap index. Note that we initialized the first tap of RFFICA for the reference channels as a pseudo-inversion of the steering vector stack for the current experiment so that we can assign 1 to the target direction and null to the interference direction:

$$\mathbf{W}_{\text{ini}}|_{\text{Ref}} = \left([\vec{D}_t\ \ \vec{D}_i]^H [\vec{D}_t\ \ \vec{D}_i] \right)^{-1} [\vec{D}_t\ \ \vec{D}_i]^H, \tag{19}$$

where $\vec{D}_t$ and $\vec{D}_i$ are steering vectors toward the target and interference direction, respectively. Because we update the initialized filter using ICA, a slight mismatch with actual DOA can be adjusted in the updating procedure. For the current experiment, we set α as 0.5 just to penalize the larger deviation from the first stage output. As a nonlinear function $g(\cdot)$, we used a polar-coordinate based tangent hyperbolic function, suitable to the super-Gaussian sources with a good convergence property [7]:

$$g(\vec{X}) = \tanh(|\vec{X}|) \exp(j \angle \vec{X}), \tag{20}$$

where $\angle \vec{X}$ represents the phase response of the complex value $\vec{X}$. To deal with the permutation and scaling, we also used the steered response of the converged first tap demixing filter as following:

$$S^i = \frac{S_i}{F_i} \cdot \left(\frac{|F_i|}{\max(|\vec{F}|)} \right)^{\gamma}, \tag{21}$$

where $i = 1, 2, \cdots, N_{mic}$ is the designated channel number, F_i is the steered response for the channel output to the target DOA [4], and $\vec{F}$ is the steered response vector to the candidate DOAs. To penalize the non-look direction in the scaling process we added the nonlinear attenuation with the normalization using the steered response. For our current experiment, we set γ as 1. The spatial filter also penalizes the non-look directional sources in each frequency bin. Note that the superiority of the subband domain approach over the time domain approach is obvious, because we can train each subband individually in a more stable manner with fewer filter coefficients.

3.2. Results

We have evaluated 18 two-speech reverberant mixing cases. With a conservative setting of 1000 iterations and 20-tap filters for each subband, the proposed approach improves SIR in the range of 10 to 29 dBC and the PESQ MOS score in the range of 0.1 to 0.6 points. Figure 1 represents contour plot

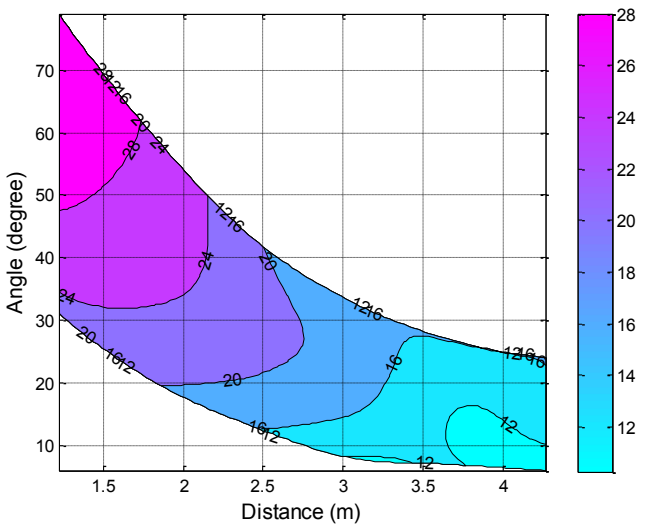

Fig. 1. Contour plot of the improvement in SIR (dBC) as a function of the distance and separation angle.

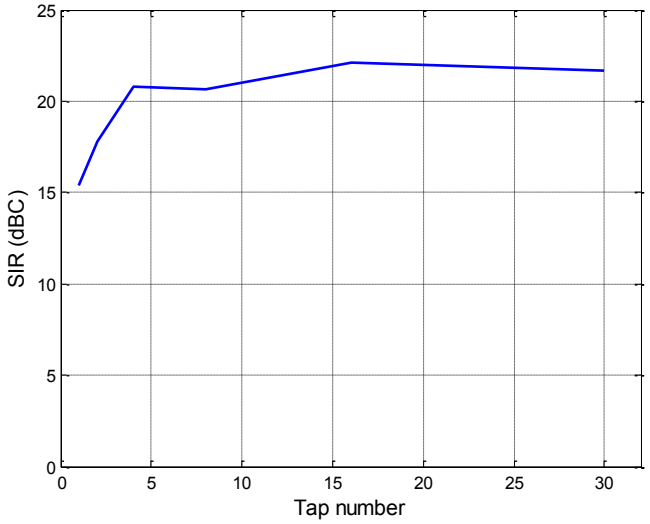

Fig. 2. SIR versus tap number.

of the improvement in SIR. Note that the contour plots have been generated by interpolating the results of 18 actual measurements. Assuming 20 dBC separation as good enough for practical purposes, we can say that the proposed algorithm is good enough for distances up to 2.7 m for two speakers at 26°, i.e. standing shoulder to shoulder. In the most difficult condition of 6° between speakers at 4.23 m distance, we can still maintain a 10 dBC SIR. Figure 2 shows the result of the dependency on tap number with the following configuration (2.44 m distance from the speech sources to microphone array and 26° angle distance between two speech sources). Note that by increasing the number of taps, the performance can also be improved as expected from the Proposition 2. Actually, with 5 taps, the SIR score can already reach the satisfactory level.

4. CONCLUSION

In this paper, we provide a theoretical justification of the huge performance improvement given by RFFICA [6]. The feedforward ICA can be explained as an direct blind adaptive method with multi-taps in subband domain and it is proven to be an equivalent implementation of the original time-domain feed-forward ICA.

5. REFERENCES

[1] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 109–116, 2005.

[2] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[3] T.-W. Lee, *Independent Component Analysis Theory and Applications*. Boston, MA: Kluwer Academic Publishers, 1998.

[4] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 666–678, 2006.

[5] H. L. V. Trees, *Optimum Array Processing Part IV of Detection, Estimation, and Modulation Theory*. New York, NY: Wiley, 2002.

[6] L.-H. Kim, I. Tashev, and A. Acero, "Reverberated speech signal separation based on regularized subband feedforward ICA and instantaneous direction of arrival," in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2010.

[7] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," in *Proc. Intern. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2002, pp. I1001–I1004.