



Audio Engineering Society Convention Paper

Presented at the 123rd Convention
2007 October 5–8 New York, NY

The papers at this Convention have been selected on the basis of a submitted abstract and extended precis that have been peer reviewed by at least two qualified anonymous reviewers. This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Convolutional blind source separation of speech signals in the low frequency bands

Maria G. Jafari¹, Mark D. Plumbley¹

¹*Centre for Digital Music, Queen Mary University of London, UK*

Correspondence should be addressed to Maria G. Jafari (maria.jafari@elec.qmul.ac.uk)

ABSTRACT

Sub-band methods are often used to address the problem of convolutional blind speech separation, as they offer the computational advantage of approximating convolutions by multiplications. The computational load, however, often remains quite high, because separation is performed on several sub-bands. In this paper, we exploit the well known fact that the high frequency content of speech signals typically conveys little information, since most of the speech power is found in frequencies up to 4kHz, and consider separation only in frequency bands below a certain threshold. We investigate the effect of changing the threshold, and find that separation performed only in the low frequencies can lead to the recovered signals being similar in quality to those extracted from all frequencies.

1. INTRODUCTION

The blind source separation (BSS) problem considered in this paper arises when two microphones record mixtures of two speech signals. Recovery of the original sources is to be performed from only the two mixtures available, and since these are convolutional in nature, the problem is often addressed in the frequency domain, where separation is carried out independently at each frequency bin. Nonetheless, frequency domain BSS (FD-BSS) is computationally expensive because source separation has to

be carried out on a large number of bins (a typical short-time Fourier transform length is 2048 point). In this paper, we take our motivation from the fact that, generally, speech signals contain little information in the frequencies above 4kHz [1], and therefore we focus on performing source separation only on the lower frequency sub-bands, while the high frequencies are set to zero. The role of the high frequency components in convolutional BSS was investigated in [2], where it was shown that, when the high frequencies are ignored, the quality of separa-

tion does not deteriorate significantly, and the recovered sources remained quite clear. Performing source separation only in the low frequencies results in reduced computational complexity of frequency domain blind source separation algorithms.

Here, we also investigate the effect of changing the cut-off frequency, i.e. the threshold above which the frequency components are set to zero. We show that the quality of the recovered signals is poor when the cut-off frequency was chosen to be below 4kHz. Also, although separation quality improved for a threshold of 4kHz, significantly better results were obtained at 4.5kHz.

2. PROBLEM FORMULATION

The problem that we seek to address is formally given by

$$x_q(n) = \sum_{p=1}^2 \sum_{l=1}^L a_{qp}(l) s_p(n-l), \quad q = 1, 2 \quad (1)$$

where $x_q(n)$ is the q -th sampled real-valued mixture, $s_p(n)$ is the p -th speech signal, $a_{qp}(l)$ denotes the impulse response from source p to sensor q , and L is the maximum length of all impulse responses [3]. Blind source separation aims to recover the sources from only the mixtures available, so that the source signals are reconstructed according to

$$y_p(n) = \sum_{q=1}^2 \sum_{l=1}^L w_{qp}(l) x_q(n-l), \quad p = 1, 2 \quad (2)$$

where $y_p(n)$ is the p -th recovered source, and $w_{qp}(l)$, are the unmixing filters which must be estimated. This is often done in the frequency domain, via the N -point short-time fourier transform of the observed signals, leading to the following mixing and separating models

$$\mathbf{X}(f, t) = \mathbf{A}(f) \mathbf{S}(f, t) \quad (3)$$

$$\mathbf{Y}(f, t) = \mathbf{W}(f) \mathbf{X}(f, t) \quad (4)$$

where $\mathbf{S}(f, t)$, and $\mathbf{X}(f, t)$ are the STFT representations of the source and mixture vectors respectively, $\mathbf{A}(f)$ and $\mathbf{W}(f)$ are the mixing and separating matrices at frequency bin f , $\mathbf{Y}(f, t)$ is the frequency domain representation of the recovered sources, and t denotes the STFT block index. Estimation of

the unmixing filters is typically performed using an independent component analysis (ICA) algorithm, operating independently on each frequency band. This, however, has the disadvantage of introducing the problem of frequency permutations, which is typically solved by clustering the frequency components of the recovered sources, often using beamforming techniques such as in [3], where the direction of arrival (DOA) of the sources are evaluated, and used to align the permutation, see [4, 3] for more details.

The FD-BSS approach can be summarised as shown in figure 1, and the algorithm that we use to perform source separation in this paper updates the unmixing filters according [5]

$$\begin{aligned} \Delta \mathbf{W}(f) &= D [\text{diag}(-\alpha_i) \\ &\quad + E \{ \phi(\mathbf{y}(f, t)) \mathbf{y}^H(f, t) \}] \mathbf{W}(f) \\ \mathbf{W}(f) &\leftarrow \mathbf{W}(f) (\mathbf{W}(f)^H \mathbf{W}(f))^{-0.5} \end{aligned} \quad (5)$$

where \mathbf{y}^H is the conjugate transpose of \mathbf{y} , $\alpha_i = E \{ y_i(f, t) \phi(y_i(f, t)) \}$, $D = \text{diag}(1/(\alpha_i - E \{ \phi'(y_i(f, t)) \}))$, and the activation function $\phi(\mathbf{y}(f, t))$ is given by

$$\phi(\mathbf{y}(f, t)) = \frac{\mathbf{y}(f, t)}{|\mathbf{y}(f, t)|}, \quad \forall |\mathbf{y}(f, t)| \neq 0 \quad (6)$$

with its derivative approximated by $\phi'(\mathbf{y}(f, t)) \approx |\mathbf{y}(f, t)|^{-1} - \mathbf{y}(f, t)^2 |\mathbf{y}(f, t)|^{-3}$ [5]. Moreover, the algorithm (5) requires that the mixtures $\mathbf{x}(f, t)$ be pre-whitened. This FD-BSS method can be found in [5], and we refer to it as MD2003.

3. BSS IN LOW FREQUENCY BANDS

Based on the observation that little information is contained in the frequencies of speech signals above a frequency of about 4kHz [1], we propose to perform source separation only on the frequency bands below a certain threshold, which we will select empirically. Thus, denoting the frequency cut-off as f_c , the following decomposition of the observed signals is considered

$$\mathbf{X}(f, t) = \mathbf{X}_{LFs}(f, t) + \mathbf{X}(f, t)_{HFs} \quad (7)$$

where $\mathbf{X}_{LFs}(f, t)$ is the STFT representation of the mixtures with the subbands corresponding to the

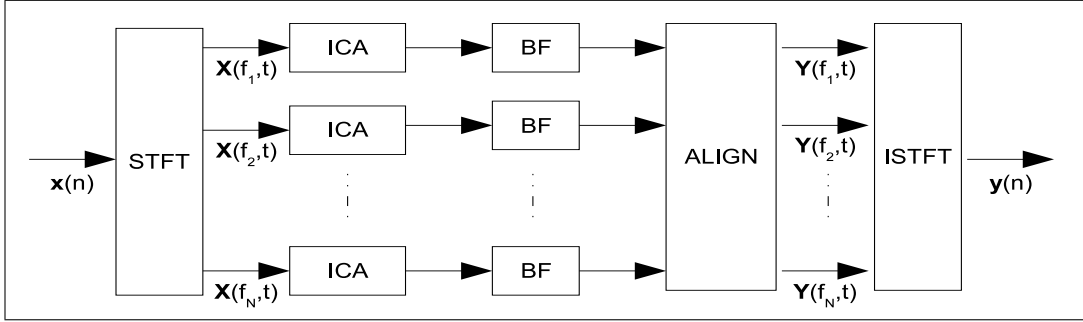


Fig. 1: Separation of all frequency bins (MD2003): $\mathbf{Y}(f, t) = \mathbf{Y}_{LFs}(f, t) + \mathbf{Y}(f, t)_{HFs}$.

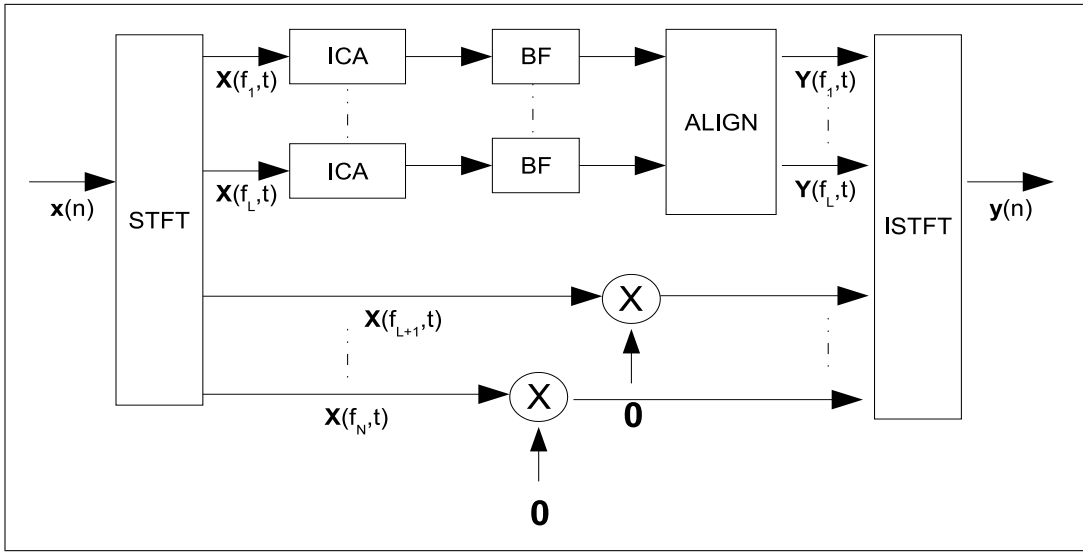


Fig. 2: Separation of low frequency bins only: $\mathbf{Y}(f, t) = \mathbf{Y}_{LFs}(f, t)$.

high frequencies ($f > f_c$) set to zero, and similarly $\mathbf{X}(f, t)_{HFs}$ has the low frequencies subbands ($f \leq f_c$) set to zero. Defining the recovered signal as $\mathbf{Y}(f, t) = \mathbf{Y}_{LFs}(f, t) + \mathbf{Y}(f, t)_{HFs}$, we perform source separation on all frequency bins, so that equation (7) holds, and on only the low frequency bins only, and thus the high frequencies are set to zero, so that (7) becomes $\mathbf{Y}(f, t) = \mathbf{Y}_{LFs}(f, t)$. This scenario is illustrated in figure 2.

4. SIMULATION RESULTS

The simulation results presented in this section were obtained from the separation of two speech signals, from two male speakers, sampled at 16kHz. The

f_c	SDR (dB)	SIR (dB)	SAR (dB)
3kHz	0.61	8.82	1.86
4kHz	1.26	9.07	2.55
5kHz	1.41	9.27	2.67
6kHz	1.48	9.37	2.72

Table 1: Signal-to-distortion (SDR), signal-to-interference (SIR), and signal-to-artifact ratios (SAR), for the four methods separating the sources signals: At cut-off varying between 3kHz and 6kHz.

sources were mixed using simulated room impulse responses, determined by the image method [6] using MGOVERN's RIR Matlab function,¹ with a room reverberation time of 320 ms. The signals were separated using the FD-BSS method in [5] (MD2003), operating on the low frequency bands only, for a cut off frequency of $f_c = \{3\text{kHz}, 4\text{kHz}, 5\text{kHz}, 6\text{kHz}\}$, and operating on all frequencies. The STFT frame length used was set to 2048 in all cases, and permutations were aligned as in [4].

The quality of the speech signals obtained with this approach was evaluated using informal listening tests, as well as the objective measures of Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR) and Signal-to-Artifacts Ratio (SAR), which measure respectively, the level of the total distortion in the estimated source, with respect to the target source, the distortion due to interfering sources, and other remaining artefacts [7]. The results presented in table 1 indicate that the quality of the separated signals is poor for a cut-off frequency of below 4kHz, while it improves above $f_c = 4\text{kHz}$, and it remains essentially unchanged as the cut-off frequency increases to 6kHz. The informal listening tests corroborated these results, and in particular it was found that the signals sounded more and more natural as f_c increased. Nonetheless, the crucial point is that these signals remained equally as intelligible as the others. Thus, it can be concluded that, for speech signals, there is little to be gained from carrying out separation in the high frequencies, which might be due to source separation methods performing worse on high frequency components, since these are generally lower in amplitude. This has the effect of reducing the computational burden of some frequency domain algorithms. The listening tests also indicated that, although separation quality improved for a threshold of 4kHz, significantly better results were obtained at 4.5kHz: the separated speech signals were found to be of similar quality to those extracted from all frequencies.

5. CONCLUSIONS

In this paper we considered applying frequency domain blind source separation methods only in frequency bands below a certain threshold, when separating convolutional mixtures of speech signals. The

effect of changing the frequency threshold was also investigated, and it was found that separation performed only in the low frequencies can lead to the recovered signals being similar in quality to those extracted from all frequencies. Moreover, the informal listening test suggested that $f_c = 4.5\text{kHz}$ is an appropriate threshold to select for separation.

6. REFERENCES

- [1] D. Balcan and J. Rosca "Independent component analysis for speech enhancement with missing TF content," in *Proc. ICA*, 2006, pp. 552–560.
- [2] M. G. Jafari, and M. D. Plumbley "The role of high frequencies in convolutional blind source separation of speech signals," To appear in *Proc. ICA*, 2007.
- [3] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. on Speech and Audio Processing*, vol. 12, pp. 530–538, 2004.
- [4] N. Mitianoudis and M. Davies, "Permutation alignment for frequency domain ICA using subspace beamforming methods," in *Proc. ICA*, 2004, pp. 669–676.
- [5] N. Mitianoudis and M. Davies, "Audio source separation of convolutional mixtures," *IEEE Trans. on Audio and Speech Processing*, vol. 11, pp. 489–497, 2003.
- [6] S. McGovern, "A model for room acoustics," Available at: <http://2pi.us/rir.html>, (2003).
- [7] C. Févotte, R. Gribonval and E. Vincent, "BSS-EVAL Toolbox User Guide," *IRISA Technical Report 1706*, April 2005. http://www.irisa.fr/metiss/bss_eval/.

¹Available from: <http://2pi.us/code/rir.m>