

The role of high frequencies in convolutive blind source separation of speech signals

Maria G. Jafari and Mark D. Plumbley *

Centre for Digital Music,
Queen Mary University of London, UK
maria.jafari@elec.qmul.ac.uk,
<http://www.elec.qmul.ac.uk>

Abstract. In this paper, we investigate the importance of the high frequencies in the problem of convolutive blind source separation (BSS) of speech signals. In particular, we focus on frequency domain blind source separation (FD-BSS), and show that when separation is performed in the low frequency bins only, the recovered signals are similar in quality to those extracted when all frequencies are taken into account. The methods are compared through informal listening tests, as well as using an objective measure.

1 Introduction

Convolutive blind source separation is often addressed in the frequency domain, through the short-time fourier transform (STFT), and source separation is performed separately at each frequency bin, thus reducing the problem to that of several instantaneous BSS problems. Although the approximation of convolutions by multiplications result in reduced computational complexity, frequency domain BSS (FD-BSS) remains computationally expensive because source separation has to be carried out on a large number of bins (a typical STFT length is 2048 point), each containing sufficient data samples for the independence assumption to hold. In addition, transforming the problem to several independent instantaneous problems, has the unwelcome side effect of introducing the problem of frequency permutations, whose solution is often quite computationally expensive [2], as it involves the clustering the frequency components of the recovered sources, using methods such as beamforming approaches, e.g. [4, 5]. These methods exploit phase information contained in the de-mixing filters identified by the source separation algorithm.

Generally, the characteristics of speech signals are such that little information is contained in the frequencies above 4kHz [14], suggesting a possible approach to BSS for speech mixtures that focuses on the lower frequencies. Motivated by this, and in order to reduce the computational load of FD-BSS algorithms, we consider here the role of high frequencies in source separation of speech signals. We show that high frequencies are not as important as low frequencies,

* This work was funded by EPSRC.

and that intelligibility is preserved even when the high frequency subbands are left unixed, and simply added back onto the separated signal. Other possible approaches would exploit existing methods that assume that high frequencies are not available, such as bandwidth extension. The structure of this paper is as follows: the basic convolutive BSS problem is described in section 2; an overview of FD-ICA is given in section 3, while the role of high frequencies is discussed in section 4. Simulation results are presented in section 5, and conclusions are drawn in section 6.

2 Problem formulation

The simplest convolutive BSS problem arises when 2 microphones record mixtures $\mathbf{x}(n)$ of 2 sampled real-valued signals, $\mathbf{s}(n)$, which in this paper are considered to be speech signals. The aim of blind source separation is then to recover the sources, from only the 2 convolutive mixtures available. Formally, the signal recorded at the q -th microphone, $x_q(n)$, is

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

where $s_p(n)$ is the p -th source signal, $a_{qp}(l)$ denotes the impulse response from source p to sensor q , and L is the maximum length of all impulse responses [2]. The source signals are then 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.

3 Frequency domain blind source separation

The convolutive audio source separation is often addressed in the frequency domain. It entails the evaluation of the N -point short-time fourier transform of the observed signals, followed by the use of instantaneous BSS, independently on each of the resulting N subbands. Thus, the mixing and separating models in (1) and (2) become, respectively

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

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

where t denotes the STFT block index.

FD-BSS has the drawback 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 [2, 4–6], where the direction of arrival (DOA) of the sources are evaluated from the beamformer directivity patterns

$$F_p(f, \theta) = \sum_{q=1}^2 W_{qp}^{\text{ICA}}(f) e^{j2\pi f d \sin \theta_p / c}, \quad p = 1, 2 \quad (5)$$

where W_{qp}^{ICA} is the ICA de-mixing filter from the q -th sensor to the p -th output, d is the spacing between two sensors, θ_p is the angle of arrival of the p -th source signal, and $c \approx 340\text{m/s}$ is the speed of sound in air. The frequency permutations are then determined by ensuring that the directivity pattern for each beamformer is approximately aligned along the frequency axis.

The BSS algorithm considered in this paper is given in [11]. It updates the unmixing filters according to

$$\begin{aligned} \Delta \mathbf{W}(f) &= D [\text{diag}(-\alpha_i) + E \{ \phi(\mathbf{y}(f, t)) \mathbf{y}^H(f, t) \}] \mathbf{W}(f) \\ \mathbf{W}(f) &\leftarrow \mathbf{W}(f) (\mathbf{W}(f)^H \mathbf{W}(f))^{-0.5} \end{aligned} \quad (6)$$

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

$$\phi(\mathbf{y}(f, t)) = \frac{\mathbf{y}(f, t)}{|\mathbf{y}(f, t)|}, \quad \forall |\mathbf{y}(f, t)| \neq 0 \quad (7)$$

and its derivative can be approximated by $\phi'(\mathbf{y}(f, t)) \approx |\mathbf{y}(f, t)|^{-1} - \mathbf{y}(f, t)^2 |\mathbf{y}(f, t)|^{-3}$ [11]. Moreover, the algorithm (6) requires that the mixtures $\mathbf{x}(f, t)$ be pre-whitened; we refer to it as MD2003.

4 The role of high frequencies

In this paper, we aim to investigate the role of the high frequencies in convolutive blind source separation of speech signals, whose characteristics are such that little information is contained in the frequencies above 4kHz [14]. Here, we consider the following decomposition of the observed signal

$$\mathbf{X}(f, t) = \mathbf{X}_{LFs}(f, t) + \mathbf{X}(f, t)_{HFs} \quad (8)$$

where $\mathbf{X}_{LFs}(f, t)$ is the STFT representation of the mixtures with the subbands corresponding to the high frequencies (above 4kHz) set to zero, and similarly $\mathbf{X}(f, t)_{HFs}$ has the low frequencies subbands (below 4kHz) set to zero. Defining the recovered signal as $\mathbf{Y}(f, t) = \mathbf{Y}_{LFs}(f, t) + \mathbf{Y}(f, t)_{HFs}$, the following four scenarios are considered, in which source separation is performed using MD2003:

1. on **all** frequency bins (MD2003): $\mathbf{Y}(f, t) = \mathbf{Y}_{LFs}(f, t) + \mathbf{Y}(f, t)_{HFs}$
2. on the **low** frequency bins only; the high frequencies are set to **zero** (LF):
 $\mathbf{Y}(f, t) = \mathbf{Y}_{LFs}(f, t)$

3. on the **low** frequency bins; the high frequency components are extracted using a beamformer $\mathbf{W}_{BF}(f)$ based on the DOAs estimated from the low frequency components (LF-BF): $\mathbf{Y}(f, t) = \mathbf{Y}_{LFs}(f, t) + \mathbf{W}_{BF}(f)\mathbf{X}(f, t)_{HFs}$
4. on the **low** frequency bins; the high frequency components are left mixed, and they are added back to the separated low frequencies prior to applying the inverse STFT (LF-BF): $\mathbf{Y}(f, t) = \mathbf{Y}_{LFs}(f, t) + \mathbf{X}(f, t)_{HFs}$

Figure 1 illustrates the four methods described above.

5 Simulation results

In this section, we consider the separation of two speech signals, from two male speakers, sampled at 16kHz. The sources were mixed using simulated room impulse responses, determined by the image method [3] using MGovern’s RIR Matlab function¹, with a room reverberation time of 160 ms. The STFT frame length used was set to 2048 in all cases. The performance of the FD-BSS method in [11] (MD2003) was compared for the four methods described in section 4, and permutations were aligned as in [4]. The low frequency bands we work with are between 0 to 4.7kHz, and high frequencies refers to frequencies above 4.7kHz.

Method	SDR (dB)	SIR (dB)	SAR (dB)	Listening Tests
MD2003 [11]	5.37	19.17	6.08	+++
LF	5.37	19.66	5.59	+
LF-BF	5.15	17.33	5.52	++
LF-HF	5.04	13.16	6.14	++++

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 all frequencies - MD2003; At low frequencies only - LF; At low frequencies; BF applied at high frequencies - LF-BF; At low frequencies; high frequencies added still mixed - LF-HF, for a cut off of 4.7kHz.

The performance of each method was evaluated using the objective criteria of Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR) and Signal-to-Artifacts Ratio (SAR), as defined in [12]. SDR, SIR and SAR 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. The evaluation criteria allows for the recovered sources to be modified by a permitted distortion, and we considered a time-invariant filter of length 512 samples, when calculating the performance measures. This length was chosen so

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

that the filter would cover the reverberation time. We obtained SDR, SIR and SAR figures for the four methods, and for all sources and microphones. The results are shown in Table 1, where the single figure was produced by averaging the criteria across all microphones and all sources.

The SDRs in Table 1 show that the total distortion for all methods is essentially the same. Distortion increases for LF-HF, due to the high frequencies not being separated, and therefore re-introducing some level of distortion. This is supported by the corresponding SIR figure for the same method, which shows that a higher level of interference from the other source is present. The values for SAR indicate that most artefacts are introduced when separation is performed on the low frequency (LF) components only, and when the high frequency components are extracted using beamforming (LF-BF). This is hardly surprising, since both methods can have quite severe effects on the data. The most interesting result is observed from the SIR figures. They show that separating only the low frequency components, and truncating the high frequency ones, has the effect of removing more interference from the undesired source signal than when working with all frequencies, while not introducing any additional distortion (SDR is unchanged), although the level of artefacts present increases. This result is rather counterintuitive, as it suggests that there is little to be gained from performing separation in the high frequencies. This might be explained by the fact that source separation methods perform worse on high frequency components, which are generally lower in amplitude; using beamforming methods to deal with the permutation problem also yields poor results due to phase ambiguity in the high frequencies [13].

Informal listening tests were performed, to corroborate the outcome of the objective criteria. They indicated that the ratios are a good guide to the audible performance. The outputs of LF were found to sound the least natural among all the recovered signals, due to the high frequencies not being present, while the sources separated with LF-HF were found to sound somehow better than the outputs of MD2003. However, the crucial point is that the outputs of all methods sounded similar in quality, suggesting that they all have similar performance. The last column in Table 1 shows a classification of the recovered sources, with the number of + indicating how good the quality of the separated signal is. In general, LF-HF gave the best results, and LF is the worst only because it is not as natural as the others. Nonetheless, the output of LF is equally as intelligible as the others.

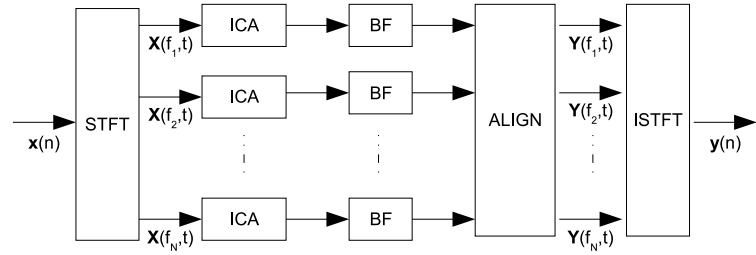
We can conclude from these results that performing separation in all subbands is not always the best approach. Especially for speech signals, it might be more advantageous to apply BSS only in the low frequencies, hence reducing, or even halving, the computational burden of some frequency domain algorithms.

6 Conclusions

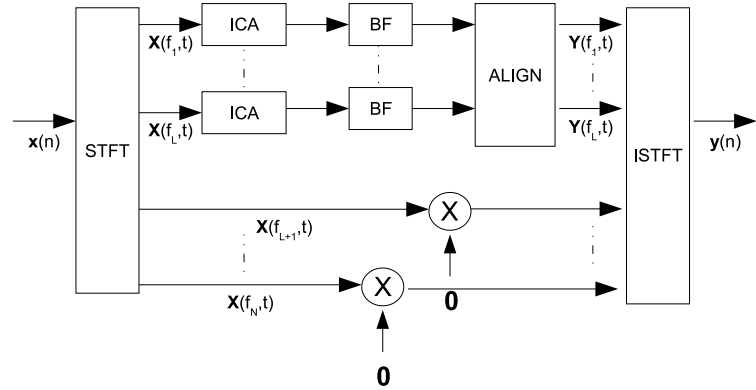
In this paper, we discussed the role of the high frequencies in frequency domain blind source separation of speech signals. We found that when the high frequencies are ignored, the separated sources remain quite clear, albeit they do not always sound very natural. Our findings were supported by objective criteria, and informal listening tests, which have suggested that it might be a good strategy to separate the mixtures in the low frequencies only, and then add on the high frequency components, without performing any processing on them. This approach may bring significant advantages in terms of reduced computational complexity.

References

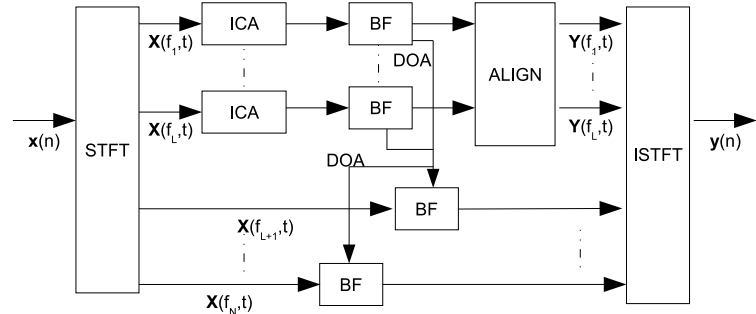
1. J.F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, pp. 2009–2025, 1998.
2. H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. on Speech and Audio Processing*, vol. 12, pp. 530–538, 2004.
3. S. McGovern, "A model for room acoustics," Available at: <http://2pi.us/rir.html>, (2003).
4. N. Mitianoudis and M. Davies, "Permutation alignment for frequency domain ICA using subspace beamforming methods," in *Proc. ICA*, 2004, pp. 669–676.
5. H. Saruwatari, S. Kurita, and K. Takeda, "Blind source separation combining frequency-domain ICA and beamforming," in *Proc. ICASSP*, 2001, vol. 5, pp. 2733–2736.
6. M. Ikram and D. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. ICASSP*, 2002, vol. 1, pp. 881–884.
7. J.-H. Lee, T.-W. Lee, H.-Y. Jung, and S.-Y. Lee, "On the efficient speech feature extraction based on independent component analysis," *Neural Processing Letters*, vol. 15, pp. 235–245, 2002.
8. S. Adballah and M. Plumbley, "Application of geometric dependency analysis to the separation of convolved mixtures," in *Proc. ICA*, 2004, pp. 22–24.
9. J.F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing*, vol. 44, pp. 3017–3030, 1996.
10. C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustic, Speech, and Signal Processing*, vol. 24, pp. 320–327, 1976.
11. N. Mitianoudis and M. Davies, "Audio source separation of convolutive mixtures," *IEEE Trans. on Audio and Speech Processing*, vol. 11, pp. 489–497, 2003.
12. 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/.
13. M. G. Jafari, S. A. Adballah, M. D. Plumbley, and M. E. Davies "Sparse coding for convolutive blind audio source separation," in *Proc. ICA*, 2006, pp. 132–139.
14. D. Balcan and J. Rosca "Independent component analysis for speech enhancement with missing TF content," in *Proc. ICA*, 2006, pp. 552–560.



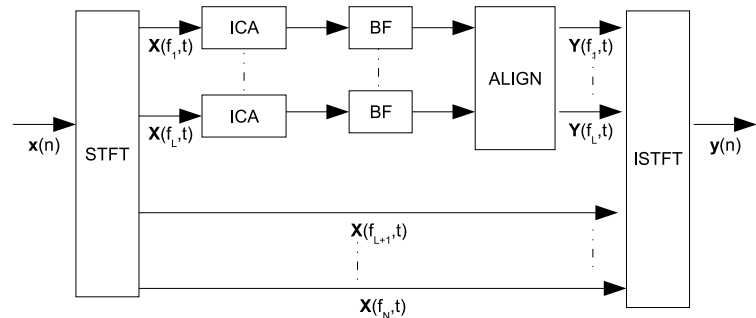
(a) Separation of all frequency bins (MD2003): $\mathbf{Y}(f, t) = \mathbf{Y}_{LFs}(f, t) + \mathbf{Y}(f, t)_{HF_s}$



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



(c) Separation of low frequency bins, with beamforming in the high frequencies (LF-BF): $\mathbf{Y}(f, t) = \mathbf{Y}_{LFs}(f, t) + \mathbf{W}_{BF}(f) \mathbf{Y}(f, t)_{HF_s}$



(d) Separation of low frequency bins. High frequency are added back without separation (LF-HF): $\mathbf{Y}(f, t) = \mathbf{Y}_{LFs}(f, t) + \mathbf{X}(f, t)_{HF_s}$

Fig. 1. Illustration of the four methods compared.