Interspeech 2018
2-6 September 2018, Hyderabad

Speech Source Separation using ICA in Constant Q Transform Domain

D. V. L. N. Dheeraj Sai, K. S. Kishor and K. Sri Rama Murty

Department of Electrical Engineering , Indian Institute of Technology, Hyderabad
ee16mtech11003@iith.ac.in, ee15mtech11002@iith.ac.in, ksrm@iith.ac.in

Abstract

In order to separate individual sources from convoluted speech mixtures, complex-domain independent component analysis (ICA) is employed on the individual frequency bins of time-frequency representations of the speech mixtures, obtained using short-time Fourier transform (STFT). The frequency components computed using STFT are separated by constant frequency difference with a constant frequency resolution. However, it is well known that the human auditory mechanism offers better resolution at lower frequencies. Hence, the perceptual quality of the extracted sources critically depends on the separation achieved in the lower frequency components. In this paper, we propose to perform source separation on the time-frequency representation computed though constant Q transform (CQT), which offers non uniform logarithmic binning in the frequency domain. Complex-domain ICA is performed on the individual bins of the CQT in order to get separated components in each frequency bin which are suitably scaled and permuted to obtain separated sources in the CQT domain. The estimated sources are obtained by applying inverse constant Q transform to the scaled and permuted sources. In comparison with the STFT based frequency domain ICA methods, there has been a consistent improvement of 3dB or more in the Signal to Interference Ratios of the extracted sources.

Index Terms: source separation, complex-valued ICA, constant Q transform, non uniform logarithmic binning.

1. Introduction

Blind Source Separation (BSS) is the task of separating individual sources from recorded mixtures without using any prior knowledge about either the source signals or the mixing parameters. The statistical independence between the sources has been exploited in independent component analysis (ICA) to achieve source separation [1]. Several measures of nongaussianity have been proposed to extract statistically independent sources [2]. In the case of instantaneous mixtures, ICA can be directly applied in the time domain to extract the sources, as a direct correspondence is maintained between respective samples in the mixtures [3]. However, in the case of convolutive mixtures, the sources signal are convolved with the impulse response of acoustic path between the source and the microphone before getting added, and are given by

$$x_i[n] = \sum_{j=1}^{N} \sum_{\tau} a_{ij}[\tau] s_j[n-\tau], \qquad \text{for } i = 1, 2, \dots M \quad (1)$$

where $x_i[n]$ denotes the i^{th} mixture signal, M denotes the number of mixture signals, $s_j[n]$ denotes j^{th} source signal, N denotes the number of source signals, and $a_{ij}[n]$ denotes the impulse response along the acoustic path between j^{th} source and i^{th} mixture. As a result, time-domain ICA cannot be readily extended to the convolutive mixtures because of every sample in the mixture signal is related to multiple samples in the source signals.

The convolutive mixture in (1) can be written as multiple instantaneous mixtures in the time-frequency (TF) domain [4] [5] as

$$\mathbf{X}[k,l] = \mathbf{A}[k]\mathbf{S}[k,l] \qquad \text{for } k = 1, 2, \dots K \qquad (2)$$

where $\mathbf{X}[k,l] = [X_1[k,l]\ X_2[k,l]\ \dots X_M[k,l]]^T$ is the vector of time-frequency representations of the mixtures in k^{th} frequency bin of the l^{th} time frame, $\mathbf{S}[k,l] = [S_1[k,l]\ S_2[k,l] \dots S_N[k,l]]^T$ is the vector of TF representations of the source signals, $\mathbf{A}[k] = A_{ij}[k]$ is matrix of k^{th} coefficients in the frequency domain representation of $a_{ij}[n]$, K is the number of frequency bins and L is the number of time frames. Notice that in this formulation the mixing matrix $\mathbf{A}[k]$ is assumed to be independent of time, and hence is valid only for stationary speakers.

Short-time Fourier transform (STFT) is typically used to compute the TF representations of the mixture signals $X_j[k,l]$ in (2) [6]. The frequency components computed using STFT are separated by constant frequency difference with a constant frequency resolution. However, it is well known that human auditory mechanism has better resolution at lower frequencies than at higher frequencies [7]. Hence, the perceptual quality of the separated sources depend on the degree of separation in the lower frequency bands, which can be achieved by choosing a TF representation with higher resolution at low frequency bands. In this paper, we propose to use constant-Q transform (CQT) to compute the time-frequency representations in (2). The CQT offers better resolution at low frequencies, and thus prevents spectral smearing in the low frequency regions [8]. The advantage of CQT is it's ability to adopt window sizes based on the frequency bin, i.e. longer analysis window for low frequency bins, and shorter analysis windows in high frequency bins [9]. As a result, the length of the analysis window need not be adjusted in the TF representations obtained through CQT. On the other hand, the performance of the STFT based methods critically depends on the size of the analysis window, which should be in the same range of the length of the impulse response of the acoustic path. The CQT has been used along with Degenerate Unmixing Estimation Technique (DUET) for separation of sources from under-determined mixtures [10].

Since all the terms in (2) are in complex domain, we need to apply complex-valued ICA [11] in order to estimate the source signals $\mathbf{S}[k,l]$, and mixing matrix $\mathbf{A}[k]$ in the frequency domain. Most of the ICA algorithms such as robust ICA [12], INFOMAX [13], complex Entropy Bound Minimization (CEBM) [14] are based on the estimation of unmixing matrix $\mathbf{W}_k$ such that the extracted sources denoted by

10.21437/Interspeech.2018-1732

$\mathbf{Y}[k,l] = [Y_1[k,l]\ \ Y_2[k,l] \ldots Y_N[k,l]]^T$ given by,

$$\mathbf{Y}[k,l] = \mathbf{W}_k^H \mathbf{X}[k,l] \tag{3}$$

are maximally independent, or equivalently maximally non-gaussian. Mutual information, which can be explicitly expressed in terms of higher order statistics, is a straight forward measure to quantify the degree of independence between the random variables [15]. However, the resulting contrast functions may not be robust and efficient. Nonlinear contrast functions, which implicitly embed the higher order statistics into the algorithm, have emerged as an alternative measure for non-gaussianity [16]. It has been proved that the extrema of the arbitrary nonlinear contrast functions coincide with the independent components. The choice of the nonlinear contrast function depends on the probability distribution of the source signal. For example, maximizing the nonlinear contrast function of the form $1/(1+z^2)$ leads to estimation of source signals that closely follow Cauchy distribution. For the case of speech separation, the choice of the contrast function should depend on the spectral characteristics of the speech in a particular frequency bin. In this paper, we present a detailed performance analysis of four different contrast functions for blind speech source separation.

Since the complex-valued ICA is applied independently on each of the frequency bins, the resulting spectral components of the estimated source signals suffer from permutation and scaling issues [17]. Several attempts have been made to address the issue of permutation by using direction of arrival of source signals [18], correlation of the envelopes of the estimated weight vectors across the frequency bins [19], and correlations between bin-wise power ratios of STFT coefficients of signals [20]. In this paper we propose to use the temporal correlations across frequency bins to correct the permutation. The within source average of the correlation coefficients, computed between pairs of frequency bins, should be higher than across-source average of the correlation coefficients. The ratio of within source to across source averages is used iteratively to arrive at the correct source ordering. The performance of the proposed CQT based method is found to be better than the conventional STFT based methods under different reverberant conditions.

The rest of the sections in this paper are organized as follows. A detailed description of the proposed algorithm is given in section 2. Section 3 presents the evaluation of the proposed algorithm and respective improvements in the performances. Conclusions are presented in section 4.

2. Complex ICA in CQT domain

2.1. Motivation

In this section, we describe the CQT approach for BSS. Human auditory mechanism shows constant Q characteristics from 500Hz to 20kHz [7]. Generally speech is phonetically more dominant at low frequencies and hence low frequency components are better resolved by the ear [21]. So, giving emphasis to the low frequency components can be of greater aid for intelligibility. In order to emphasize only the low frequency components more, non uniform binning needs to be carried out which cannot be done with STFT approach. Hence we consider the approach of CQT, through which non uniform binning can be achieved. Fig. 1 shows the sequence of steps that have been adopted in the proposed algorithm for BSS.

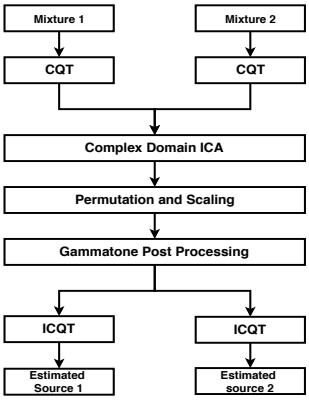

Figure 1: *Block diagram of the proposed algorithm*

2.2. The CQT TF representation

The STFT of the signal $x[n]$ using a window $w_f[n]$ is defined as follows [22],

$$X[k] = \sum_{n=0}^{N_f-1} w_f[n]x[n]e^{-j\frac{2\pi}{N_f}kn} \tag{4}$$

where $X[k]$ is k^{th} spectral component, N_f point discrete Fourier transform (DFT) and digital frequency $2\pi k/N_f$. Thus there exists a linear spacing of TF bins. In order to provide a non uniform binning and an improved low frequency resolution, we adopt CQT. The CQT of the signal $x[n]$ is defined as follows:

$$X[k] = \frac{1}{N_f[k]} \sum_{n=0}^{N_f[k]-1} w_f[k,n]x[n]e^{-j\frac{2\pi Q}{N_f[k]}n} \tag{5}$$

The digital frequency in this case turns out to be $Q\frac{2\pi}{N_f[k]}$. If the k^{th} spectral component is denoted by f_k, it can be formulated as,

$$f_k = (2^{\frac{1}{b}})^k f_{min} \tag{6}$$

where number of bins per octave is denoted by b. The parameter Q is defined as the ratio of central frequency to the bandwidth which can be formulated as follows,

$$Q = \frac{f_k}{\Delta f_k} = \frac{f_k}{f_{k+1} - f_k} = \frac{1}{2^{\frac{1}{b}} - 1} \tag{7}$$

But this Q transform is non-invertible. We need an invertible Q transform in order to convert estimated sources back to time domain. Hence to have an invertible CQT, we have considered the approach of CQT developed in the context of non stationary Gabor transform (CQ-NSGT) where the inversion of the transform has been achieved by the introduction of dual frames, where frame is an orthonormal basis generalization [23]. For this purpose, a non stationary Gabor filter bank with constant Q factor in every bin is developed whose bandwidth (BW) is defined as follows,

$$BW = \alpha * f_k + \gamma \tag{8}$$

where

$$\alpha = 2^{\frac{1}{b}} - 2^{\frac{-1}{b}} \tag{9}$$

where the parameter γ is the BW offset. If the offset is greater than zero, the BW of the filter increases which leads to an improved resolution towards low frequencies. Since complex ICA requires a TF representation, except for the highest frequency channel, rest of the channels are sampled at the same sub sampling rate.

In order to ensure perfect reconstruction, the DC frequency and higher frequency components are also considered in CQT-NSGT at the same computational cost unlike the conventional CQT approach.

2.3. Complex ICA

ICA is applied on mixtures which uses CQT to get the estimated sources. In this paper, we have adopted a computationally efficient approach where nonlinear estimators are utilized to exploit the higher order statistics [24]. The function that has been considered for optimization is given by,

$$J_G(\mathbf{w}) = E[G(|\mathbf{w}^H\mathbf{X}|^2)] \tag{10}$$

where $\mathbf{w}$ is the unmixing weight vector to be estimated. We have adopted four different nonlinearities *G(z)* such as $\sqrt{0.1+z}$, $\frac{1}{(0.1+z)}$, *tanh(z)* and $z^2/2$.

By considering the constraint $\|\mathbf{w}\|^2 = 1$, the optimal points of above function are computed as shown below,

$$\bigtriangledown E[G(|\mathbf{w}^H\mathbf{X}|^2)] - \beta \bigtriangledown E[(|\mathbf{w}^H\mathbf{X}|^2)] = 0 \tag{11}$$

The second term can be approximated using a jacobian matrix and then the above equation is solved using newton's method. On further simplification, a more compact version of the iterative weight update equation is obtained which is given as,

$$\begin{aligned}\mathbf{w}^+ = E[\mathbf{X}(\mathbf{w}^H\mathbf{X})^* g(|\mathbf{w}^H\mathbf{X}|^2)] - E[g(|\mathbf{w}^H\mathbf{X}|^2) \\ +(|\mathbf{w}^H\mathbf{X}|^2)(g^{'}(|\mathbf{w}^H\mathbf{X}|^2))]\mathbf{w}\end{aligned} \tag{12}$$

where $g()$ is derivative of $G()$ and $g^{'}()$ is derivative of $g()$. The updated weights are then normalized. In order to solve for the scaling issue, we need to adjust estimated unmixing weight matrices $\mathbf{W_k}$ such that they have determinant of unity [4].

$$\mathbf{W_k}^{new} = \mathbf{W_k}^{original}.|\mathbf{W_k}^{original}|^{\frac{-1}{N}} \tag{13}$$

After obtaining the estimated source frequency bins, they are suitably permuted in order to align them in the corresponding source channels. In order to permute the bins, envelope correlation coefficient approach has been considered. To depermute, we define an envelope correlation coefficient [25] which is formulated as,

$$\rho_{ij}[k,l] = \frac{\sum_{q=1}^{Q} v_i[k,q]v_j[l,q]}{\sqrt{\sum_{q=1}^{Q} v_i^2[k,q]}\sqrt{\sum_{q=1}^{Q} v_j^2[l,q]}} \tag{14}$$

where k and l are two adjacent frequency bins and $v_i[k,:] = |Y_i[k,:]|$ (: is a Matlab convention adopted to denote all the elements in a row) is the amplitude envelope of the i^{th} extracted source output of complex ICA. For instance, considering a 2 microphone and 2 speaker scenario, the alignment can be done by defining a constant ζ such that ,

$$\zeta[k,l] = \frac{\rho_{11}[k,l] + \rho_{22}[k,l]}{\rho_{12}[k,l] + \rho_{21}[k,l]} \tag{15}$$

If $\zeta[k,l]$ is greater than one, it implies that the adjacent bins along the same channel are correlated more and need not be permuted and vice verse. After resolving the permutation and scaling issues, the extracted sources are post processed before converting back to time domain using Gammatone filterbank [26] in order to reduce the perceptual artifact and provide a better enhancement which is described in the following section 2.4.

2.4. Post processing using Gammatone filter bank

In order to reduce the perceptual artifact in the extracted sources, channel weighting technique have been applied [27]. Initially binary masks are created by applying MAP criterion to each of the TF locations of the extracted sources. If we consider the two extracted sources as $y_1[n]$) and $y_2[n]$ and β is a threshold, then the binary mask $\mu(k,l)$ is constructed as follows,

$$\mu[k,l] = \begin{cases} 1+\varepsilon & if \quad |Y_1[k,l]| \geqslant \beta * |Y_2[k,l]| \\ \varepsilon & if \quad |Y_1[k,l]| < \beta * |Y_2[k,l]| \end{cases} \tag{16}$$

where ε corresponds to the noise floor in order to reduce the effects of musical noise [28]. After construction of the binary mask, channel weighting technique has been applied to convert the binary mask to continuous mask. The channel weighting coefficients at a TF location can be obtained by computing the ratio of power at a TF bin after binary mask application to the original input power at the m^{th} gammatone frequency channel as shown below. The extracted sources are then enhanced using the above obtained channel weighting coefficients as follows,

$$\hat{w}[m,l] = \frac{\sum_{k=0}^{\frac{K}{2}} \mu[k,l]\,|Y_a[k,l]H[k,m]|^2}{\sum_{k=0}^{\frac{K}{2}} |Y_a[k,l]H[k,m]|^2} \tag{17}$$

where $Y_a[k,l]$ is the spectral average from both the extracted sources, K denotes the number of spectral points and $H[k,m]$ is the frequency response of the m^{th} gammatone filter. Enhanced spectrum $\hat{Y}[k,l]$ is given by

$$\hat{Y}[k,l] = \sum_{m=0}^{M^{'}-1} (\sqrt{\hat{w}[m,l]}Y_a[k,l]H[k,m]) \tag{18}$$

where total number of gammatone filter banks is denoted by $M^{'}$ and $\hat{Y}[k,l]$ is the extracted source after post processing in TF domain. The above same procedure is repeated for binary mask $1-\mu[k,l]$ to get another source. The extracted sources can be obtained by converting them from time frequency domain back to time domain. In order to reconstruct them back to time domain, Inverse constant Q transform (ICQT) is applied to the estimated source frequency bins.

3. Performance Evaluation

In order to evaluate the performance of the proposed algorithm, we have considered a two microphone and two

Table 1: *Comparing the SDR and SIR values of extracted sources from speech + speech mixture with reverberation of 0.2s.*

Metric	Source	PARAFAC algorithm	$1/(0.1+z)$		$\tanh z$		$\sqrt{0.1+z}$		$z^2/2$	
			STFT	CQT	STFT	CQT	STFT	CQT	STFT	CQT
SIR(dB)	male	13.07	16.4	15.4	12.9	13.17	10.91	15.04	10.91	15.02
	female	7.91	14.03	14.33	15.76	14.54	15.37	13.72	15.16	13.53
SDR(dB)	male	-18.31	2.95	2.97	0.06	0.02	-0.11	2.96	-0.19	1.43
	female	-22.4	1.47	1.46	2.99	2.93	3.01	1.43	2.94	2.94

Table 2: *Comparing the SDR and SIR values of extracted sources from speech + speech mixture with reverberation of 0.4s.*

Metric	Source	PARAFAC algorithm	$1/(0.1+z)$		$\tanh z$		$\sqrt{0.1+z}$		$z^2/2$	
			STFT	CQT	STFT	CQT	STFT	CQT	STFT	CQT
SIR(dB)	male	11.07	14.58	14.25	5.53	7.28	5.68	14.29	14.62	14.01
	female	0.39	7.71	9.46	13.24	12.415	13.15	8.79	8.08	8.66
SDR(dB)	male	-18.8	-4.6	-4.34	-9.26	-9.27	-9.26	-4.33	-4.56	-4.48
	female	-27.4	-6.68	-6.72	-4.31	-4.08	-4.29	-6.72	-6.63	-6.63

speaker scenario. We have adopted the cases of speech-speech mixtures generated from artificial room impulse responses (RIRs) [29] with a reverbaration of 0.2s and 0.4s. All the original sources and mic recordings were obtained from `http://www.telecom.tuc.gr/~nikos/BSS_Nikos.html`. The original sources have been sampled at 16 kHz. The Sources have been extracted using the four non linearities that have been mentioned. For the purpose of evaluation, source extraction has been done using both STFT and CQT TF representations. Signal to Distortion Ratios (SDR) and Signal to Interference ratios (SIR) have been adopted as metrics in order to compare the TF representations and the values have been tabulated in Table 1 and Table 2.

For extracting the sources using STFT time frequency representation, we have experimented over various window lengths to get the best performance and a window length of 0.128s i.e. 2048 samples has been chosen for both reverberation times of 200 *ms* and 400 *ms*. The overlap factor has been taken as 0.75 and number of FFT points were taken to be equal to the window length. The parameters of constant Q transform have to be fixed beforehand, and for the current evaluation the minimum frequency to be analyzed has been fixed at 100Hz and the maximum frequency to be analyzed has been fixed at half the sampling frequency i.e. 8 kHz. The number of bins per octave were taken to be 100 and windowing is done using hamming window. The same parameters have been considered for both the cases of 200 *ms* and 400 *ms*. The CQT has been computed using CQT matlab toolbox [30] [31]. SIRs have been evaluated using BSS_EVAL toolbox [32]. The SIRs and SDRs for the sources extracted using PARAFAC algorithm [33] have also been computed using BSS_EVAL toolbox by comparing against the original sources. In order to perform post processing using Gammatone filter bank, the total number of filters have been taken to be 64. To compensate for the trade off between the intelligibility and quality of the extracted sources, the threshold for constructing the binary mask has been fixed at $\beta = 1$.

We have evaluated the performance of the proposed approach against various source separation algorithms other than PARAFAC such as PARARA [34] which exploits the non stationarity property of the sources through least square optimization , Time frequency BSS method [20] in which extraction is done by clustering of the mixtures in TF domain using angle between a reference and sample vector. The proposed approach has also shown a good performance under high reverberant conditions. The results have been tabulated in Table 3.We can observe that emphasizing the low frequencies using CQT results in significant enhancement in the extracted sources in comparison with STFT based methods which is clearly evident in the improved SIRs and SDRs. A part of improvement in SDR can also be attributed to gammatone post processing as the perceptual artifact has been reduced. In order to reduce the effects of musical noise, a small amount of noise floor is added to the binary mask and then processed [28]. As a result of this, speech intelligibility can be preserved. On the other hand, addition of greater magnitude of noise floor degrades the speech quality. Hence an optimal value of noise floor has to be chosen to preserve both the intelligibility as well as quality of the extracted sources. For the current evaluation it has been set to $\epsilon = 0.003$. The results can be found at `https://sites.google.com/a/iith.ac.in/bss_demo/`

Table 3: *Comparison of SIR(dB) values against various source separation algorithms for speech + speech mixtures having reverbaration time of 200 ms and 400 ms.*

Algorithm	speaker	200 *ms*	400 *ms*
PARARA [34]	male	8.49	1.77
	female	11.08	6.2
TF domain bss [20]	male	16.32	9.09
	female	13.62	5.63
PARAFAC [33]	male	13.07	11.07
	female	7.91	0.39
CQT approach	male	15.4	14.25
	female	14.33	9.46

4. Conclusions

The main objective of this paper is to get a better enhanced version of the extracted sources from recorded convolutive mixtures through constant Q TF representation. Non uniform binning of the mixtures has been achieved through this approach which aids in the emphasis of the corresponding low frequency components. The performance has been tested over different reverberant conditions and over various bss algorithms. From the experimental results, we find that the approach has exhibited a good performance also under high reverberant conditions.

5. References

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2004, vol. 46.

[2] J.-M. Lee, C. Yoo, and I.-B. Lee, "Statistical process monitoring with independent component analysis," *Journal of Process Control*, vol. 14, no. 5, pp. 467–485, 2004.

[3] Z. Koldovskỳ and P. Tichavskỳ, "Time-domain blind audio source separation using advanced ica methods," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[4] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[5] C. Févotte, A. Debiolles, and C. Doncarli, "Blind source separation of fir convolutive mixtures: Application to speech signals," in *ISCA Tutorial and Research Workshop on Non-Linear Speech Processing*, 2003.

[6] N. Yang, M. Usman, X. He, M. A. Jan, and L. Zhang, "Time-frequency filter bank: A simple approach for audio and music separation," *IEEE Access*, vol. 5, pp. 27 114–27 125, 2017.

[7] C. Schörkhuber and A. Klapuri, "Constant-q transform toolbox for music processing," in *7th Sound and Music Computing Conference, Barcelona, Spain*, 2010, pp. 3–64.

[8] A. Nagathil and R. Martin, "Evaluation of spectral transforms for music signal analysis," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[9] T. Fillon and J. Prado, "A flexible multi-resolution time-frequency analysis framework for audio signals," in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*. IEEE, 2012, pp. 1124–1129.

[10] Z. Rafii and B. Pardo, "Degenerate unmixing estimation technique using the constant q transform," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 217–220.

[11] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[12] V. Zarzoso and P. Comon, "Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size," *IEEE Transactions on Neural Networks*, vol. 21, no. 2, pp. 248–261, 2010.

[13] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[14] X.-L. Li and T. Adali, "Complex independent component analysis by entropy bound minimization," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 7, pp. 1417–1430, 2010.

[15] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[16] E. Bingham and A. Hyvärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *International journal of neural systems*, vol. 10, no. 01, pp. 1–8, 2000.

[17] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*. Springer, 2007, vol. 615.

[18] 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 transactions on speech and audio processing*, vol. 12, no. 5, pp. 530–538, 2004.

[19] B.-R. Chen, H.-Y. Lee, and Y.-W. Liu, "Unmixing convolutive mixtures by exploiting amplitude co-modulation: Methods and evaluation on mandarin speech recordings," *Proc. Interspeech 2017*, pp. 1934–1937, 2017.

[20] V. G. Reju, S. N. Koh, and Y. Soon, "Underdetermined convolutive blind source separation via time–frequency masking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 101–116, 2010.

[21] D.-C. Balcan and J. Rosca, "Independent component analysis for speech enhancement with missing tf content," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2006, pp. 552–560.

[22] J. C. Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[23] G. A. Velasco, N. Holighaus, M. Dörfler, and T. Grill, "Constructing an invertible constant-q transform with non-stationary gabor frames," *Proceedings of DAFX11, Paris*, pp. 93–99, 2011.

[24] E. Bingham and A. Hyvarinen, "Ica of complex valued signals: a fast and robust deflationary algorithm," in *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, vol. 3. IEEE, 2000, pp. 357–362.

[25] R. Mazur and A. Mertins, "Solving the permutation problem in convolutive blind source separation," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 512–519.

[26] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[27] C. Kim and K. K. Chin, "Sound source separation algorithm using phase difference and angle distribution modeling near the target," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[28] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79.*, vol. 4. IEEE, 1979, pp. 208–211.

[29] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[30] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.

[31] N. Holighaus, M. Dörfler, G. A. Velasco, and T. Grill, "A framework for invertible, real-time constant-q transforms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 775–785, 2013.

[32] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[33] D. Nion, K. N. Mokios, N. D. Sidiropoulos, and A. Potamianos, "Batch and adaptive parafac-based blind separation of convolutive speech mixtures," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1193–1207, 2010.

[34] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.