International Journal of Speech Technology (2020) 23:451–457
https://doi.org/10.1007/s10772-020-09706-x

A study on unsupervised monaural reverberant speech separation

R. Hemavathi[1] · **R. Kumaraswamy**[1]

Received: 17 November 2019 / Accepted: 10 April 2020 / Published online: 9 May 2020

Abstract

Separating individual source signals is a challenging task in musical and multitalker source separation. This work studies unsupervised monaural (co-channel) speech separation (UCSS) in reverberant environment. UCSS is the problem of separating the individual speakers from multispeaker speech without using any training data and with minimum information regarding mixing condition and sources. In this paper, state-of-art UCSS algorithms based on auditory and statistical approaches are evaluated for reverberant speech mixtures and results are discussed. This work also proposes to use multiresolution cochleagram and Constant Q Transform (CQT) spectrogram feature with two-dimensional Non-negative matrix factorization. Results show that proposed algorithm with CQT spectrogram feature gave an improvement of 1.986 and 1.262 in terms of speech intelligibility and 0.296 db and 0.561 db in terms of signal to interference ratio compared to state-of-art statistical and auditory approach respectively at T60 of 0.610s.

Keywords Unsupervised speech separation · Speech intelligibility · Reverberant environment · Monaural recordings · Non-negative matrix factorization

1 Introduction

Multispeaker environment refers to scenarios like conferences and meetings, where the intended speech is degraded with interference speech. The speech signal collected in multispeaker environment is termed as multispeaker (multitalker) speech. Humans have the ability to listen to the intended speaker's speech in multitalker environment (Cherry 1953; Mesgarani and Chang 2012). Studies show speech/speaker recognition systems give degraded performance in realistic environments in presence of interference speech and reverberation (Rennie et al. 2010). By being able to separate the desired speech from multispeaker speech better performance can be obtained in recognition systems.

Many algorithms are proposed to separate the intended speech from multitalker speech. Based on number of observation (recording) required, the algorithms are classified under single channel and multi-channel speech separation algorithms. Multichannel speech separation (MCSS) algorithms require recordings from two or more microphones (Yegnanarayana et al. 2009). State-of-art multichannel speech separation algorithms are based on principles of Independent Component Analysis (ICA) (Hyvarinen 1999; Chien and Hsieh 2012), beamforming (Li et al. 2014; Madhu and Martin 2011; Saruwatari et al. 2006) and signal processing (Yegnanarayana et al. 2005; Swamy et al. 2007). Statistical approaches based on ICA and its variants give excellent performance for artificial speech mixtures but give degraded performance in realistic scenarios. Algorithms based on signal processing exploit delay and speaker specific features like pitch and give excellent performance in reverberant and noisy conditions. Multispeaker speech, y_m collected using $m = 1, 2 \dots M$ number of sensors (microphones) using $n = 1, 2 \dots N$ number of sources (speakers), in noisy and reverberant environment is given by,

$$y_m(t) = \sum_{n=1}^{N} \sum_{l=0}^{L-1} a_{mn}(l) s_n(t-l) + v(t) \tag{1}$$

where $a_{mn}(l)$ is the impulse response of the path l from speaker n to sensor m, where $l = 0, 1, 2 \dots L$ and $v(t)$ is additional noise (Yegnanarayana et al. 2009).

✉ R. Hemavathi
hemavathir@sit.ac.in

R. Kumaraswamy
hyrkswamy@sit.ac.in

1 Department of Electronics and Communication Engineering, Siddaganga Institute of Technology (Affiliated to Visvesvaraya Technological University, Belagavi), Tumakuru 572103, India

This work focuses on single-channel speech separation, also referred to as monaural or cochannel speech separation (Molla and Hirose 2007). Separating the desired speech from interference speech using co-channel speech is challenging task as, both desired and unwanted signal have same characteristics and unlike MCSS, spatial information is not available. Cochannel speech separation problem is addressed in both supervised (Wang and Chen 2018; Krishna and Ramaswamy 2017; Smaragdis 2007) and unsupervised (Gao et al. 2011; Hu and Wang 2013; Gao et al. 2013) perspective. Supervised speech separation algorithms addressed the problem of separation of desired speech signal from interfering non-speech noise (Wang and Wang 2013) and speech signal (Wang and Wang 2019). The algorithms are also extended to reverberant environment in recent works (Delfarah and Wang 2017). Supervised speech separation is accomplished using models based on, Deep Neural Networks (Wang and Wang 2013), stack of DNNs (Zhang and Wang 2016) and combining DNN with Recurrent Neural Networks (RNNs) (Huang et al. 2015). These algorithms require clean utterance to be available and knowledge of interfering signal. In realistic scenarios these conditions are hard to meet. Hence, this paper focuses on unsupervised single channel source separation.

This paper addresses separation of multi-speaker speech collected in reverberant environment. Main contribution of this paper are:

1. This paper studies unsupervised cochannel speech separation (UCSS) in reverberant environment. The State-of-art UCSS based on non-negative matrix factorization and auditory analysis is studied for reverberant speech. As per our knowledge this is the first work to address separation of desired speech from interference speech in reverberant condition with monaural recording in unsupervised perspective.
2. This paper proposes to use Multi-resolution cochleagram and CQT spectrogram features with two dimensional NMF for speech separation in reverberant environment
3. Proposed algorithms showed to increase the intelligibility of speech compared to state of art statistical and auditory approaches and also successively suppress the interference speech.

Rest of the paper is organized as follows, Sect. 2 gives the review of unsupervised monaural speech separation, Sect. 3 discusses the proposed methodology, Sect. 4 gives experimental details and in Sects. 5 and 6 the results are discussed and the work is concluded.

2 Related work

This section gives the overview of work done in unsupervised co-channel speech separation. State-of-art UCSS are based on Non-negative Matrix Factorization (NMF), multipitch estimation, Auditory Scene Analysis (ASA), and Empirical Mode Decomposition (EMD) techniques.

NMF based techniques decomposes the non-negative matrix input into the product of basis vectors H matrix and encoding matrix W,

$$Y = HW \tag{2}$$

Multitalker speech separation using basis matrix doesn't give good performance as there will be overlap between subspaces. Hence, separation is accomplished using sparse NMF (SNMF) in Schmidt and Olsson (2006). SNMF optimizes the cost function in (3)

$$E = \|Y - \bar{H}W\|_F^2 + \lambda \sum_{ij} W_{ij} \;\; s.t. \;\; H, W \geq 0 \tag{3}$$

where $\bar{H}$ is the column wise dictionary matrix, λ controls the degree of sparsity. Itakuro saito NMF was proposed by using cochleagram representation of mixed speech in Gao et al. (2013). This paper showed good separation for music mixtures and for speech mixtures it gave relatively degraded performance.

CASA based systems perform speech separation by exploiting features like offsets, onsets and periodicity from TF representation of input mixed speech. Using these features, in primitive stage individual segments are formed and in secondary stage the individual segments are grouped into foreground and background streams. Finally, separated speaker's speech is obtained by resynthesizing the streams. Based on fact that the performance of the speech separation do not depend only on Signal to Noise Ratio (SNR), an approach was proposed by combining CASA with objective quality assessment of speech (OQAS) in Li et al. (2006). The TF representation of input speech was obtained using auditory filter bank and features like autocorrelation, envelope, dominant pitch, and cross channel correlation are extracted. Based on these features, segments are formed are grouped into desired and interfering speech. These works focused on speech separation for voiced speech. An unsupervised method for cochannel speech separation which aims at separating both voiced and unvoiced speech is presented in Hu and Wang (2013). This algorithms made use of tandem algorithm and performed separation in two stages for voiced and unvoiced speech.

Pitch detection approaches for source separation are classified into frequency and time domain approaches. Frequency domain approaches performs separation by locating harmonic peaks and time domain approaches includes autocorrelation and average magnitude difference function for pitch detection. UCSS system based on harmonic enhancement of desired speech and suppression of interference speech is presented in Morgan et al. (1995). This paper uses maximum likelihood pitch detector. To improve speaker identification in cochannel speech a UCSS system is presented in Shao and Wang (2003). These algorithms also assume instantaneous mixing, non-overlapping speech and aims at separating dominant pitch from the multitalker speech.

Empirical Mode Decomposition (EMD) is an approach for analysing nonlinear and non stationary signals. The EMD decomposes mixed speech adaptively into Intrinsic Mode Functions (IMFs). UCSS is performed by using EMD and multipitch information in Prasanna Kumar and Kumaraswamy (2017). UCSS based on the EMD and variable regularized two-dimensional sparse non-negative matrix factorization (v-SNMF2D) is proposed in Gao et al. (2011). Instead of processing the mixed signal directly, it proposes to utilize the IMFs as the new set of observations. It benefits conventional SNMF2D in terms of improved accuracy in resolving spectral bases and temporal codes which were previously not possible by using SNMF2D.

UCSS algorithms proposed so far performed speech separation assuming instantaneous mixing condition as shown below

$$y(t) = x_1(t) + x_2(t) \tag{4}$$

where $x_1(t)$ is intended speaker's speech and $x_2(t)$ is interference speech and $y(t)$ is speech collected at sensor, which is additive mixture of $x_1(t)$ and $x_2(t)$.

Algorithms based on pitch tracking and CASA, performed well for separation of voiced speech and for unvoiced speech separation still remains challenging. EMD when combined with NMF based techniques, performed better for non-overlapping speech. But for overlapping speech, it gave degraded performance as spectral characteristics of speech overlap. In spite of many works done in UCSS, the algorithm's performance in realistic environment is not studied.

3 Proposed methodology

In this work co-channel speech is modelled as shown below,

$$y(t) = \sum_{n=1}^{N} \sum_{l=0}^{L-1} a_{nl} s_n(t-l) \tag{5}$$

$y(t)$ is the cochannel speech collected at microphone, S_n is the n^{th} speaker's speech arriving at time t and a_{nl} is impulse response of the path l from speaker n to microphone, where $l = 0, 1, \ldots L$

The main challenge in UCSS is that there will be minimal information regarding speaking and mixing condition, desired and noise signal have same characteristics and no spatial information could be exploited. Hence choosing a suitable feature plays a crucial role in performance of speech separation. This paper proposes to use Multiresolution cochleagram and Constant Q transform spectrogram feature with 2 dimensional Non-negative matrix Factorization with expectation maximization and multiple gradient descent learning algorithms. These are used to compute basis and encoding matrix from which individual speech signals are reconstructed.

3.1 Feature extraction

1. Cochleagram: Initially mixed speech signal $y(t)$ is passed through the gammatone filterbank (GFB) with the impulse response,

$$g(f,t) = \begin{cases} \mathrm{t}^{\mathrm{h}-1} e^{-2\pi v t}, & t \geq 0 \\ 0, & else \end{cases} \tag{6}$$

   where h is the filter order, v is rectangular bandwidth and center frequency f. GFB output $y(c, t)$ is given by

$$y(c,t) = y(t) * g(f_c, t) \tag{7}$$

   Each filter's output is split into time frames and their TF spectra are taken to get the cochleagram representation.
2. Multi-resolution Cochleagram: Multiresolution cochleagram was first proposed in Chen et al. (2014). MRCG can be obtained from the cochleagram using following steps. Given input mixed speech, initially with 20 ms frame size and 10 ms frame shift, a 64 channel cochleagram, $CG1$ is derived. Again with frame size 200ms, cochleagram $CG2$ is derived. By averaging $CG1$ across window 11X11, $CG3$ is obtained and $CG4$ is obtained by averaging $CG1$ across window 23X23. $CG1$ to $CG4$ are concatenated to obtain MRCG feature.
3. Constant Q Transform Spectrogram: Constant Q Transform focuses on maintaining ratio of central frequency, f_c to frequency resolution f_r constant. CQT spectrogram is obtained by stacking columns of CQT from time segments. Motivated by fact that CQT showed clear spectral variations and was used for multisource pitch extraction for music signals (Smaragdis 2009), this paper proposes to use CQT spectrogram for reverberant speech separation.

3.2 UCSS using 2 dimensional NMF

This section gives the overview of 2 dimensional NMF. One dimensional NMF given in (2) gave poorer performance for overlapping basis vector, hence NMF was extended to 2 dimension. When Time-Frequency (TF) representation of mixed speech is given as input to NMF2D, it is decomposed into product of two non-negative matrices H and W as shown below,

$$|Y|^{\cdot 2} \approx \sum_{\tau,\phi} \overset{\downarrow\phi}{\mathbf{H}^{\tau}} \overset{\rightarrow\tau}{\mathbf{W}^{\phi}} \tag{8}$$

where $\downarrow \phi$ denotes down shift of elements in matrix by ϕ rows and $\rightarrow \tau$ denotes right shift of elements in matrix by τ columns

Algorithm 1: NMF using Expectation maximization

Input: Time Frequency representation of multitalker speech signal $|S_{tf}|^2$
Output: Encoding matrix H, weight matrix W
1. Initialize H^{τ} , W^{ϕ}
2. Initialize the cost value $-logp(C_k|\theta_k)$
where, $C_k = [c_{k,1}, c_{k,2}, ...c_{k,t}]$ is complex Gaussian distribution whose components are mutually independent and $\theta_k = \{H_k^{\tau}, W_k^{\phi}\}$
3. for n=1 : niter
$E - step$: Compute posterior power of C_k
$M - step$: Untill convergence update $H_{f,t}^{\tau}$ and $W_{f,t}^{\phi}$ using Expectation maximization algorithm as in [7].
end
4. Stopping criteria: $(cost(n-1) - cost(n))/cost(n) < \psi$, where $\psi = 10^{-6}$

Algorithm 2: NMF using Multiplicative Gradient Descent learning

Input: Time Frequency representation of multitalker speech signal $|Y|^2$
Output: Encoding matrix H, weight matrix W
1. Initialize H^{τ} , W^{ϕ}
2. Initialize the cost value $-logp(Y|H, W)$, under independent and IID noise assumption
3. for n=1 : niter
compute $Z = \sum_{\tau} \sum_{\phi} H_{f-\phi}^{\tau} W_{t-\tau}^{\phi}$
update H^{τ} and W^{ϕ} using MGD algorithm as in [7]
end
4. stopping criteria: $(cost(n-1) - cos(n))/cost(n) < \psi$, where $\psi = 10^{-6}$

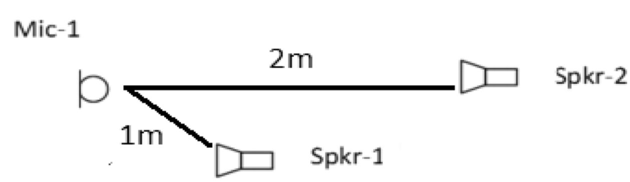

Fig. 1 Experimental setup for data collection

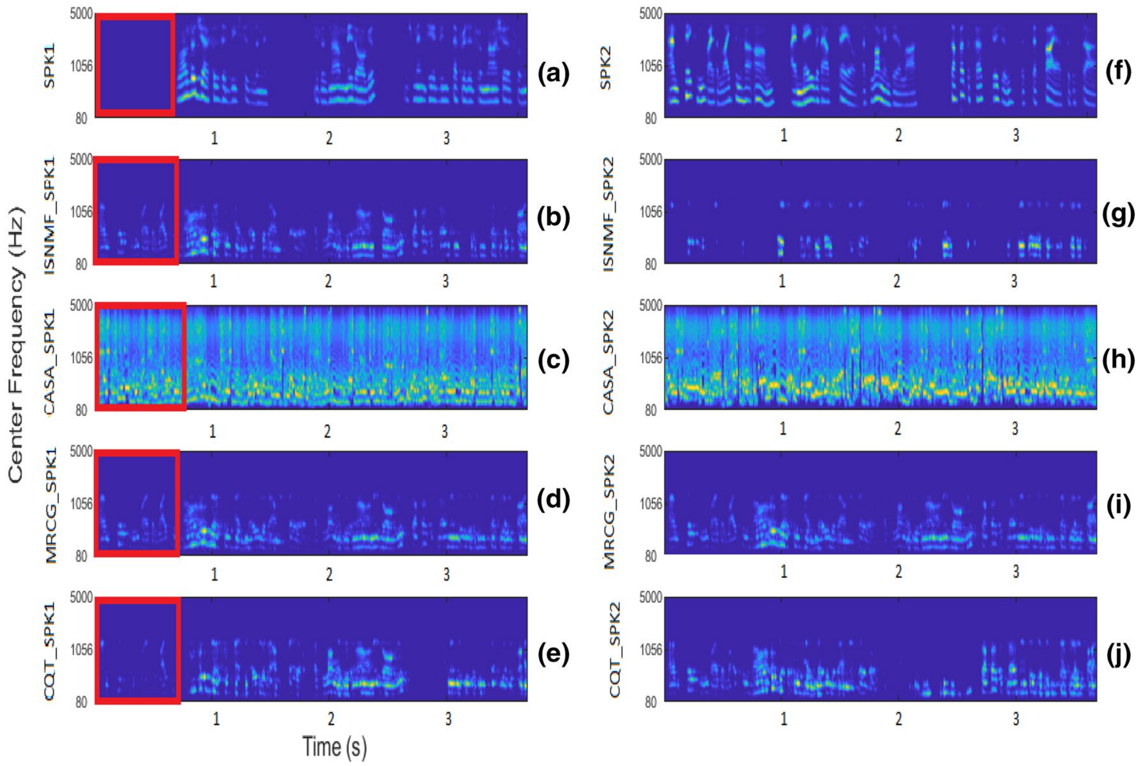

Fig. 2 Cochleagram plots of **a** and **f** clean speech signal from speaker 1 and speaker 2 respectively **b–e** separated speaker 1 speech from ISNMF, CASA, MRCG-EM, CQT-EM respectively **g–j** separated speaker 2 speech from ISNMF, CASA, MRCG-EM, CQT-EM respectively

Table 1 STOI and SIR scores of separated desired and interference speech signal from different speech separation algorithms at various T60

Speech separation algorithms	SIR						STOI					
	T60 = 0.160		T60 = 0.360		T60 = 0.610		T60 = 0.160		T60 = 0.360		T60 = 0.610	
	SPK1	SPK2	SPK1	SPK2	SPK1	SPK2	SPK1	SPK2	SPK1	SPK2	SPK1	SPK2
ISNMF	2.24	2.09	1.062	0.829	0.117	− 0.683	0.325	0.309	0.3429	0.335	0.376	0.370
CASA	1.52	− 0.050	1.37	− 0.353	0.8413	− 0.5322	0.2077	0.150	0.173	0.058	0.1097	− 0.12
MRCG-EM	3.326	1.700	0.238	− 0.884	1.424	− 0.251	0.390	0.388	0.356	0.349	0.331	0.329
MRCG-MGD	1.201	1.041	− 0.062	− 1.902	− 0.026	− 0.52	0.356	0.349	0.331	0.329	0.312	0.307
CQT-EM	5.942	4.475	4.968	3.466	2.103	− 0.23	0.6909	0.448	0.6820	0.434	0.6706	0.365
CQT-MGD	4.583	3.784	3.031	2.405	1.490	− 1.14	0.687	0.440	0.675	0.389	0.66	0.312

To decompose the given input matrix, H and W are learnt using two algorithms Expectation maximization and MGD. Initially the TF representation of mixed speech, $|Y|^2$ is obtained by passing through filter bank. This is given as input for NMF2D algorithm. NMF decomposes the given non-negative input $|Y|^2$ into H, basis matrix and W, encoding matrix. The algorithm of EM-NMF and MGD-NMF are given in Algorithm1 and Algorithm2 respectively.

4 Experimental setup

For experiments the two speaker reverberant speech mixtures are generated by taking clean speech signals from TIMIT database and impulse responses from RIR generator (Hadad et al. 2014). Desired speech is placed at a distance of 1m and interfering speaker at a distance of 2m from microphone at 45 and 90 degrees respectively as shown in Fig. 1. The speech signal are mixed at various Signal to

Reverberation Ratios (SRR) and all speech signals are sampled at a rate of 16 kHz with the average duration of 2 to 4 s.

For both EM-NMF and MGD-NMF parameters are τ ans ϕ are set to vary from 0:1, 0:3, number of iteration is fixed to 50, number of Channels are 128 and length of gammatone filter is 128ms. For cochleagram feature, frequency range is [50 5000], window length is 20ms and frame shift is 10ms. For Constant Q Spectrogram frequency resolution is 2, frequency range is [55 5000] and time resolution is 25. Multiresolution Cochleagram is obtained from cochleagram as shown in Sect. 3.1.

For convenience the Multiresolution cochleagram and Constant Q transform spectrogram feature used with expectation maximization learning is termed as MRCG-EM and CQT-EM and with multiple gradient descent learning is termed as MRCG-MGD and CQT-MGD respectively. In this work, Unsupervised speech separation algorithms based on auditory and statistical models are studied in reverberant environment. UCASA (Hu and Wang 2013) is chosen in auditory approach, it gave better performance compared to supervised algorithms. ISNMF (Gao et al. 2013) which is computationally efficient and give better performance for speech mixtures is chosen in statistical approach.

For evaluation of all algorithms, 200 reverberant speech mixtures were generated with various T60s (0.160 s, 0.360 s and 0.610 s). Short time objective intelligibility (STOI) (Taal et al. 2010) and Signal to Interference Ratio (SIR) are used to evaluate performance of the algorithms and results are tabulated in Table 1. STOI gives intelligibility measure and SIR is measure of interference suppression.

5 Results and discussions

Proposed algorithms are compared with state of art unsupervised monaural speech separation algorithm based on ASA and statistical approach. Figure 2 shows that separated speech resultant from *ISNMF* and *CASA* retained speech from interference speech. Whereas, proposed CQT-EM and MRCG-EM suppressed the interfering speech (which can be seen in red box). Desired speech is better re-constructed in CQT/MRCG -EM-NMF compared to UCASA and ISNMF. Table 1 shows, statistical approach, *ISNMF* gave better performance compared to Auditory approach, *UCASA*. This is because speech resultant from *UCASA* was missing some desired speech information and artefacts introduced due to Ideal Binary Masking decreased the speech intelligibility. Proposed algorithms gave better performance compared to *ISNMF* and *UCASA* algorithms, because of features used for speech separation. Proposed UCSS algorithms with MRCG feature gives additional contextual information by exploiting local and global information using multi-resolution feature extraction. Hence, it performed better compared to gammatone feature, *ISNMF*. CQT gave excellent performance compared to all algorithms by decreasing the reverberation effect and reconstructing both desired and unwanted signals. For non overlapping speech good separation is obtained, but for overlapping speech the unwanted speech is suppressed but not completly removed. EM-NMF gave better performance compared to MGD-NMF, for both features, this is because MGD algorithm may not yield better H and W as it tends to be trapped in local minima. EM algorithm prevents zero entries and in MGD zero coefficients are considered and these remain invariant during updates. As the stationary point is the limit in MGD, it is hard to determine when it attains a fixed point solution with zero entries (Fevotte et al. 2009). In terms of computation time, MRCG-EM consumed more time due to multiresolution property, followed by CASA, ISNMF and CQT-EM. The challenge in UCSS is absence of spatial information and no information regarding desired or interference speaker. Hence, feature selection plays a crucial role in UCSS. Proposed CQT-EM gave better performance in terms of improvement in intelligibility, suppression of interference and fast computation.

6 Conclusion

This paper studies speech separation in reverberant environment using unsupervised learning approaches. Unsupervised single channel speech separation is most challenging problem in source separation as there will be no information regarding speakers or mixing condition and unlike MCSS there will be no spatial information. This is the first work to study co-channel speech separation in reverberation environment in unsupervised perspective. This paper shows that unsupervised approach can perform better for overlapping speech in reverberant conditions too, if proper feature and better learning algorithms are used. The proposed algorithms gave better performance compared to state of art unsupervised speech separation algorithms in terms of speech intelligibility and interference suppression. Further various features can be explored to obtain better performance and UCSS study can be extended to noisy environments.

Funding Funding was provided by Department of science and Technology, Government of India (Grant No. SR/WOS-A/ET-69/2016).

References

Chen, J., Wang, Y., & Wang, D. (2014). A feature study for classification-based speech separation at very low signal-to-noise ratio. In *IEEE International conference on acoustics, speech and signal processing ICASSP 2014, Florence, Italy, May 4–9* (pp. 7039–7043). https://doi.org/10.1109/ICASSP.2014.6854965.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229.

Chien, J. T., & Hsieh, H. L. (2012). Convex divergence ica for blind source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 302–313. https://doi.org/10.1109/TASL.2011.2161080.

Delfarah, M., & Wang, D. (2017). Features for masking-based monaural speech separation in reverberant conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(5), 1085–1094.

Fevotte, C., Bertin, N., & Durrieu, J. (2009). Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Computation*, *21*(3), 793–830. https://doi.org/10.1162/neco.2008.04-08-771.

Gao, B., Woo, W. L., & Dlay, S. S. (2011). Single-channel source separation using emd-subband variable regularized sparse features. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 961–976. https://doi.org/10.1109/TASL.2010.2072500.

Gao, B., Woo, W. L., & Dlay, S. S. (2013). Unsupervised single-channel separation of nonstationary signals using gammatone filterbank and itakura saito nonnegative matrix two-dimensional factorizations. *IEEE Transactions on Circuits and Systems I: Regular Papers*, *60*(3), 662–675. https://doi.org/10.1109/TCSI.2012.2215735.

Hadad, E., Heese, F., Vary, P., & Gannot, S. (2014). Multichannel audio database in various acoustic environments. In *2014 14th international workshop on acoustic signal enhancement (IWAENC), Juan-les-Pins, France* (pp. 313–317). https://doi.org/10.1109/IWAENC.2014.6954309.

Huang, P., Kim, M., Hasegawa-Johnson, M., & Smaragdis, P. (2015). Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(12), 2136–2147. https://doi.org/10.1109/TASLP.2015.2468583.

Hu, K., & Wang, D. (2013). An unsupervised approach to cochannel speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(1), 122–131. https://doi.org/10.1109/TASL.2012.2215591.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626–634. https://doi.org/10.1109/72.761722.

Krishna, P. K. M., & Ramaswamy, K. (2017). Single channel speech separation based on empirical mode decomposition and hilbert transform. *IET Signal Processing*, *11*(5), 579–586. https://doi.org/10.1049/iet-spr.2016.0450.

Li, P., Guan, Y., Xu, B., & Liu, W. (2006). Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(6), 2014–2023. https://doi.org/10.1109/TASL.2006.883258.

Li, W., Wang, L., Zhou, Y., Dines, J., Doss, M. M., Bourlard, H., et al. (2014). Feature mapping of multiple beamformed sources for robust overlapping speech recognition using a microphone array. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 2244–2255.

Madhu, N., & Martin, R. (2011). A versatile framework for speaker separation using a model-based speaker localization approach. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(7), 1900–1912.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(5), 233–236.

Molla, M. K. I., & Hirose, K. (2007). Single-mixture audio source separation by subspace decomposition of hilbert spectrum. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(3), 893–900. https://doi.org/10.1109/TASL.2006.885254.

Morgan, D. P., George, E. B., Lee, L. T., & Kay, S. M. (1995). Cochannel speaker separation. In *1995 international conference on acoustics, speech, and signal processing, Detroit, Michigan* (Vol. 1, pp. 828–831). https://doi.org/10.1109/ICASSP.1995.479822.

Prasanna Kumar, M. K., & Kumaraswamy, R. (2017). Single-channel speech separation using empirical mode decomposition and multi pitch information with estimation of number of speakers. *International Journal of Speech Technology*, *20*(1), 109–125. https://doi.org/10.1007/s10772-016-9392-y.

Rennie, S. J., Hershey, J. R., & Olsen, P. A. (2010). Single-channel multitalker speech recognition. *IEEE Signal Processing Magazine*, *27*(6), 66–80.

Saruwatari, H., Kawamura, T., Nishikawa, T., Lee, A., & Shikano, K. (2006). Blind source separation based on a fast-convergence algorithm combining ICA and beamforming. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(2), 666–678. https://doi.org/10.1109/TSA.2005.855832.

Schmidt, M. N., & Olsson, R. K. (2006). Single-channel speech separation using sparse non-negative matrix factorization. In *ISCA international conference on spoken language proceesing, INTERSPEECH, Pittsburgh, Pennsylvania*.

Shao, Y., & Wang, D. (2003). Co-channel speaker identification using usable speech extraction based on multi-pitch tracking. In ıtIEEE international conference on acoustics, speech, and signal processing, Hong Kong, China (Vol. 2, pp. II–205–8). https://doi.org/10.1109/ICASSP.2003.1202330.

Smaragdis, P. (2007). Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(1), 1–12. https://doi.org/10.1109/TASL.2006.876726.

Smaragdis, P. (2009). Relative-pitch tracking of multiple arbitary sounds. *The Journal of the Acoustical Society of America*, *125*(5), 3406–3413.

Swamy, R. K., Murty, K. S. R., & Yegnanarayana, B. (2007). Determining number of speakers from multispeaker speech signals using excitation source information. *IEEE Signal Processing Letters*, *14*(7), 481–484. https://doi.org/10.1109/LSP.2006.891333.

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *IEEE international conference on acoustics, speech and signal processing, Dallas, Texas, USA* (pp. 4214–4217). https://doi.org/10.1109/ICASSP.2010.5495701.

Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1702–1726. https://doi.org/10.1109/TASLP.2018.2842159.

Wang, Y., & Wang, D. (2013). Towards scaling up classification-based speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(7), 1381–1390. https://doi.org/10.1109/TASL.2013.2250961.

Wang, Z., & Wang, D. (2019). Combining spectral and spatial features for deep learning based blind speaker separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(2), 457–468.

Yegnanarayana, B., Swamy, R. K., & Prasanna, S. R. M. (2005). Separation of multispeaker speech using excitation information. NOLISP-2005, Barcelona, Spain (pp. 11–18).

Yegnanarayana, B., Swamy, R. K., & Murty, K. S. R. (2009). Determining mixing parameters from multispeaker data using speech-specific information. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(6), 1196–1207. https://doi.org/10.1109/TASL.2009.2016230.

Zhang, X., & Wang, D. (2016). A deep ensemble learning method for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(5), 967–977. https://doi.org/10.1109/TASLP.2016.2536478.