INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Time-frequency masking for blind source separation with preserved spatial cues

Shadi Pirhosseinloo[1], *Kostas Kokkinakis*[2]

[1]Department of Electrical Engineering and Computer Science, University of Kansas, USA
[2]Department of Speech-Language-Hearing, University of Kansas, USA
shadi@ku.edu, kokkinak@ku.edu

Abstract

In this paper, we address the problem of speech source separation by relying on time-frequency binary masks to segregate binaural mixtures. We describe an algorithm which can tackle reverberant mixtures and can extract the original sources while preserving their original spatial locations. The performance of the proposed algorithm is evaluated objectively and subjectively, by assessing the estimated interaural time differences versus their theoretical values and by testing for localization acuity in normal-hearing listeners for different spatial locations in a reverberant room. Experimental results indicate that the proposed algorithm is capable of preserving the spatial information of the recovered source signals while keeping the signal-to-distortion and signal-to-interference ratios high.

Index Terms: Time-frequency mask, spatial cues, dereverberation, degree of separation, localization.

1. Introduction

In realistic listening situations, human listeners excel at hearing out a specific sound of interest (target) from amongst a mixture of other interfering sounds. Inspired by this robust performance, research has been devoted to build speech separation systems that incorporate the known principles of auditory perception [1]. According to this theory, listeners perform segregation of a binaural mixture in a two-stage process. In the first stage, the acoustic input is analyzed to form time-frequency (T-F) segments, while in the second stage, listeners group sound elements based on whether those originate from the same locations (likely to come from a common source) or spatially distributed locations (likely to come from two different sources) [2]. To facilitate this latter process, listeners may rely on the differences in the overall intensity or level of the signals received at the two ears, known as interaural level differences (ILDs) and the different arrival times of signals at each ear due to the spatial separation of the two ears, known as interaural time differences (ITDs) [2, 3]. The techniques of separating individual sound sources from a mixture are known as blind source separation and computational auditory scene analysis (CASA). Both fields, have become popular in the recent decades and a number of methods have emerged from the study of this problem, most of which perform well for certain types of sources, such as speech (e.g., see [4]–[9]). Speech is sparsely distributed in the time-frequency domain and even in challenging listening environments which may consist of multiple competing speakers, speech streams remain intelligible. This is mainly due to the fact that speech energy is concentrated in isolated regions in time and frequency. Binary time-frequency masks exploit this underlying sparsity and disjointness of speech spectra in their short-time-frequency representations by creating a mask that only preserves the spectro-temporal regions where the target is dominant [5]. In this paper, we investigate a widely used T-F masking algorithm called degenerate unmixing estimation technique (DUET) [8, 9] for source separation in reverberant settings. Since the performance of the DUET is known to degrade significantly in reverberation, we implement a pre-processing stage for signal dereverberation based on interaural coherence. A cue preservation method implemented as a post-processing stage is shown to preserve level and timing cues in the extracted sources. The efficiency of the proposed cue preservation stage is evaluated in a challenging scenario, while the preservation of binaural cues is measured using a realistic localization test.

2. Algorithm Formulation

2.1. Interaural coherence (IC) stage

In this section, we describe the dereverberation stage based on interaural coherence (IC) which is employed as a pre-processing step for speech enhancement before separating sources. First, the two reverberant mixture signals recorded from the left and the right channels $x_l(n)$ and $x_r(n)$ are transformed to the time-frequency domain by using the STFT which produces the complex valued spectra $x_l(\tau,\omega)$ and $x_r(\tau,\omega)$ with τ as time frame and ω as frequency band. From the time-frequency representation of the left channel mixture and the right channel mixture, the IC is estimated using the normalized cross-correlation function calculated as (e.g., see [10]–[13]).

$$IC_{lr}(\tau,\omega) = \frac{|\Phi_{lr}(\tau,\omega)|}{\sqrt{\Phi_{ll}(\tau,\omega)\Phi_{rr}(\tau,\omega)}} \tag{1}$$

where the $\Phi_{ll}(\tau,\omega), \Phi_{rr}(\tau,\omega)$ and $\Phi_{lr}(\tau,\omega)$ are the exponentially weighted short-term auto-correlation and cross-correlation functions defined as:

$$\Phi_{ll}(\tau,\omega) = \alpha\ \Phi_{ll}(\tau-1,\omega) + (1-\alpha)\ |x_l(\tau,\omega)|^2 \tag{2}$$

$$\Phi_{rr}(\tau,\omega) = \alpha\ \Phi_{rr}(\tau-1,\omega) + (1-\alpha)\ |x_r(\tau,\omega)|^2 \tag{3}$$

$$\Phi_{lr}(\tau,\omega) = \alpha\ \Phi_{lr}(\tau-1,\omega) + (1-\alpha)\ x_r(\tau,\omega)x_l^*(\tau,\omega) \tag{4}$$

where $0 \leqslant \alpha \leqslant 1$ denotes a smoothing factor. The IC describes the coherence of the left channel signal and the right channel signal which has a range of [0,1], where 1 indicates that both signals x_l and x_r are perfectly coherent. Consequently, a binary mask $\mathrm{M_{IC}}(\tau,\omega)$ is derived from the estimated interaural coherence as:

$$\mathrm{M_{IC}}(\tau,\omega) = \begin{cases} 1, & \text{if } IC_{lr}(\tau,\omega) > \mathrm{threshold}(\omega) \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where the threshold value is determined adaptively for each frequency band as $\mathrm{threshold}(\omega) = \max\{0.8, \mathrm{Q_3(IC_{lr})}\}$, where $\mathrm{Q_3}$ is the 3^{rd} quartile for each frequency band. Fig. 1 illustrates

http://dx.doi.org/10.21437/Interspeech.2017-66

the construction of the binary mask based on the histogram of the interaural coherence. The derived binary mask detects the bins where reverberant energy is dominant and retains the bins with IC close to 1. The estimated binary mask is applied to both channels of $x_l(\tau,\omega)$ and $x_r(\tau,\omega)$ producing the processed (dereverberated) mixtures. The dereverberated signals $\hat{x}_l(n)$ and $\hat{x}_r(n)$ are reconstructed using the inverse STFT and the dereverberant mixtures are then processed with the source separation algorithm described in the following section.

2.2. Cue preservation (CP) stage

In this section, we analyze the DUET algorithm for j sources in a two-microphone system configuration [8, 9]. The windowed Fourier transforms of the left $\hat{x}_l(\tau,\omega)$ and right $\hat{x}_r(\tau,\omega)$ dereverberated mixtures can be written as:

$$\begin{bmatrix}\hat{x}_l(\tau,\omega)\\ \hat{x}_r(\tau,\omega)\end{bmatrix} = \begin{bmatrix}1 & \dots & 1\\ a_1 e^{-i\omega\delta_1} & \dots & a_N e^{-i\omega\delta_N}\end{bmatrix}\begin{bmatrix}\hat{s}_1(\tau,\omega)\\ \vdots \\ \hat{s}_N(\tau,\omega)\end{bmatrix} \tag{6}$$

where N is the number of sources and a and δ represent the mixing parameters for each source for every different time-frequency point (τ,ω). The DUET separates the sources by clustering different time-frequency points based on their interaural parameters. The method essentially constructs a histogram of interaural parameters, with points weighted by their respective energy, and then selects each prominent peak in the histogram as the interaural parameters of each source. It then creates a mask for each source that retains only time-frequency points with interaural parameters near the selected peak [8, 9]. The mixing parameters for each time-frequency point are calculated based on the mixtures as follows:

$$\tilde{a}(\tau,\omega) = \left|\frac{\hat{x}_r(\tau,\omega)}{\hat{x}_l(\tau,\omega)}\right| \tag{7}$$

$$\tilde{\delta}(\tau,\omega) = -\frac{1}{\omega}\angle\left(\frac{\hat{x}_r(\tau,\omega)}{\hat{x}_l(\tau,\omega)}\right) \tag{8}$$

Instead of using the attenuation parameter $\tilde{a}(\tau,\omega)$, the symmetric attenuation $\tilde{\alpha}(\tau,\omega)$ is often utilized as shown below:

$$\tilde{\alpha}(\tau,\omega) = \left|\frac{\hat{x}_r(\tau,\omega)}{\hat{x}_l(\tau,\omega)}\right| - \left|\frac{\hat{x}_l(\tau,\omega)}{\hat{x}_r(\tau,\omega)}\right| \tag{9}$$

After obtaining $\tilde{\alpha}(\tau,\omega)$ and $\tilde{\delta}(\tau,\omega)$, a two-dimensional smoothed weighted histogram is calculated. The peak centers of the histogram represent the mixing parameters. Next, the time-frequency binary masks are reconstructed based on the mixing parameters and are applied to the mixtures to calculate the original source estimates. The DUET algorithm is capable of recovering the original signals blindly from two mixtures. However, this algorithm does not preserve the necessary spatial cues of the original source signals. In order to overcome this problem, a binaural cue preservation strategy that can be realized as an additional post-processing step added to the original DUET is described below. The peaks are located based on the DUET smoothed histogram and the peak centers determine the mixing parameters of $(\tilde{\alpha}_j, \tilde{\delta}_j)$ of the j^{th} source [8, 9]. The attenuation parameter is calculated based on the following equation:

$$\tilde{a}_j = \frac{\tilde{\alpha}_j + \sqrt{\tilde{\alpha}_j^2 + 4}}{2} \tag{10}$$

Assuming that only one source is active so that the sources are disjoint, after obtaining both $\tilde{a}_j$ and $\tilde{\delta}_j$ for each separated source, we can recover the spatial cues corresponding to that source in the time-frequency domain according to the following equation, which defines the DUET with cue preservation (DUET-CP) algorithm:

$$\begin{bmatrix}\tilde{s}_{j_L}(\tau,\omega)\\ \tilde{s}_{j_R}(\tau,\omega)\end{bmatrix} = \begin{bmatrix}1\\ \tilde{a}_j e^{-i\omega\tilde{\delta}_j}\end{bmatrix}\tilde{s}_j(\tau,\omega) \tag{11}$$

where $\tilde{s}_j(\tau,\omega)$ is the j^{th} separated signal defined as the output of the DUET and $\tilde{s}_{j_L}(\tau,\omega)$ and $\tilde{s}_{j_R}(\tau,\omega)$ are the time-frequency representations of the j^{th} cue preserved signals corresponding to the left and right channel, respectively. By comparing Eq. (6) with Eq. (11), it becomes apparent that the mixing parameters $\tilde{a}_j$ and $\tilde{\delta}_j$ are equivalent to the interaural level differences and the interaural time differences of the spatially separated original sources.

3. Experimental Results

The performance of the proposed algorithm encompassing the IC and CP stages (IC-DUET-CP) was systematically evaluated according to three different performance outcomes: (1) degree of separation, (2) degree of dereverberation, and (3) preservation of binaural cues. To measure the degree of separation, we utilized the popular metrics of the signal-to-distortion-ratio (SDR) and signal-to-interference-ratio (SIR) [14]. To compute the amount of speech dereverberation achieved, we used the speech to reverberation modulation energy ratio (SRMR) [15] and the segmental signal-to-reverberation ratio (segSRR) [16, 17, 18]. To assess the effectiveness of the algorithm in retaining the necessary spatial cues of the output signal estimates, we compared the correlation coefficient between the theoretical and estimated interaural time differences. Additionally, we performed localization listening tests with normal-hearing listeners and measured the subjective sound identification responses using the root-mean-square localization error (RMSLE).

3.1. Speech material

The algorithm was evaluated on a test set of speech signals comprised of a single randomly selected male spoken sentence. A female interferer was used as the masker with a root-mean-square value equal to the target source, such that the input SNR was equal to 0 dB. The duration of each speech signal was approximately 3 s. All signals were recorded at a sampling rate of 22,050 Hz. To generate the speech test stimuli, we used sentences from the IEEE database, which consists of phonetically balanced sentences, with each sentence being composed of approximately 7 to 12 words [19]. All signals had the same onset and were normalized to their maximum amplitude before convolving with the HRTFs. Office head-related impulse responses measured in the University of Oldenburg were used to simulate a reverberant listening condition with RT_{60} = 0.3 s [20]. For each listening scenario, a total of seven different azimuthal sound source locations were calculated for sound sources located 1 m away from the center of the listener in the azimuthal plane for every angle from -90^0 left of the listener to +90^0 to the right of the listener in 30^0 increments. In all cases, the interference source was placed directly in front of the listener at 0^0 and the target was allowed to virtually rotate around the listener in the presence of competing speech.

Table 1: *SDR and SIR values averaged over 70 mixtures.*

	SDR (dB)		SIR (dB)	
	ANE	REV	ANE	REV
Mixtures	3.07	-0.56	3.07	1.85
IC		-0.81		1.71
DUET	10.66	1.90	16.70	8.13
DUET-CP	10.66	1.90	16.70	8.13
IC-DUET-CP		2.40		8.90

3.2. Reverberation suppression with SRMR and SegSRR

In order to evaluate the performance of the reverberation suppression stage, we used two metrics, the speech to reverberation modulation energy ratio (SRMR) [15] and the segmental signal-to-reverberation ratio (segSRR) [16, 17, 18]. For the SRMR evaluation, the processed signal is passed through a 23-channel gammatone filterbank and the temporal envelope of each filter output is calculated using the Hilbert transform. Second, the extracted envelopes are multiplied by a 256-ms Hamming window and then for each critical band, the modulation spectral energy is calculated. Next, the modulation frequency bins are grouped into eight bands. Finally, the ratio of the average modulation energy for the first four bands over the average modulation energy of the last four bands is calculated as SRMR [15]. For each channel the SRMR measurements are calculated separately and averaged over all mixtures. The ΔSRMR shows the effect of processing and is expressed as:

$$\Delta\text{SRMR} = \text{SRMR}_{\text{processed}} - \text{SRMR}_{\text{reverberant}} \tag{12}$$

The segSRR estimates the energy of the direct signal compared to the reverberant energy which is equivalent to the signal-to-noise ratio (SNR) when reverberation is considered as noise [17, 18]. Therefore, the segSRR for each frame m is calculated as:

$$\text{segSRR(m)} = 10\log_{10}\left[\frac{\sum_{n=mR}^{mR+N-1} s_d^2(n)}{\sum_{n=mR}^{mR+N-1}(s_d(n)-\bar{s}(n))^2}\right] \tag{13}$$

where $s_d(n)$ is the direct signal, $\bar{s}(n)$ is the reverberant or the processed signal. The R and N values are the frame rate in samples and the total number of signal samples respectively. Finally, the segSRR is calculated as the average of segSRR(m) over all non-silence frames and the improvement of SRR is calculated as [17]:

$$\Delta\text{segSRR} = \text{segSRR}_{\text{processed}} - \text{segSRR}_{\text{reverberant}} \tag{14}$$

where the $\text{SRR}_{\text{processed}}$ metric was calculated after the two signals were processed through the dereverberation stage and $\text{SRR}_{\text{reverberant}}$ was due to the unprocessed signals [17, 18].

3.3. Separation performance with SDR and SIR

In order to measure the separation performance, the SDR and SIR criteria were used [14]. The SDR calculates the ratio of the energy in the original signal to the summation of the energy of interference, artifacts and distortion, while SIR calculates the ratio of the target energy to the interferer energy, defined as [14]:

$$\text{SDR} = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}+e_{noise}+e_{artif}||^2} \tag{15}$$

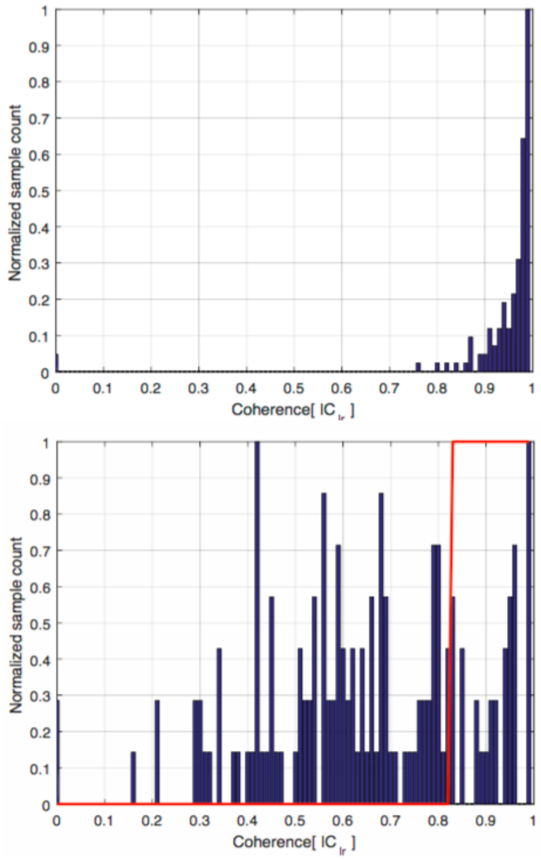

Figure 1: *The IC histogram of the 700 Hz frequency channel for an anechoic (top panel) and a reverberant signal (bottom panel) and the binary mask (red line).*

$$\text{SIR} = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2} \tag{16}$$

3.4. Localization error (RMSLE)

To evaluate the cue preservation efficiency of the proposed algorithm, the theoretical and estimated ITDs were compared. The theoretical ITD was derived based on the peak location of the cross-correlation between left and right HRTFs. The ITD was taken to be the lag at which the largest peak occurred in the cross-correlation. The ITD of the cue preserved outputs, was calculated based on the peak location of the cross-correlation between the left and right channel of the output signals. The similarity between the theoretical and experimental ITD, was assessed with the Pearson's correlation coefficient. The proposed algorithm illustrates a high correlation coefficient of $\rho = 0.97$ between the theoretical and estimated ITDs. To measure localization accuracy, six undergraduate students from the University of Kansas with American English as their first language were recruited for course credit. All listeners gave informed consent prior to testing. All listeners recruited had normal hearing and reported normal cognitive function. The listeners were presented with IEEE sentences processed in 3 different experimental conditions: (1) clean signal (target only), (2) DUET and (3) IC-DUET-CP through headphones. Each target sentence was presented in random order from any of the seven different locations for a total

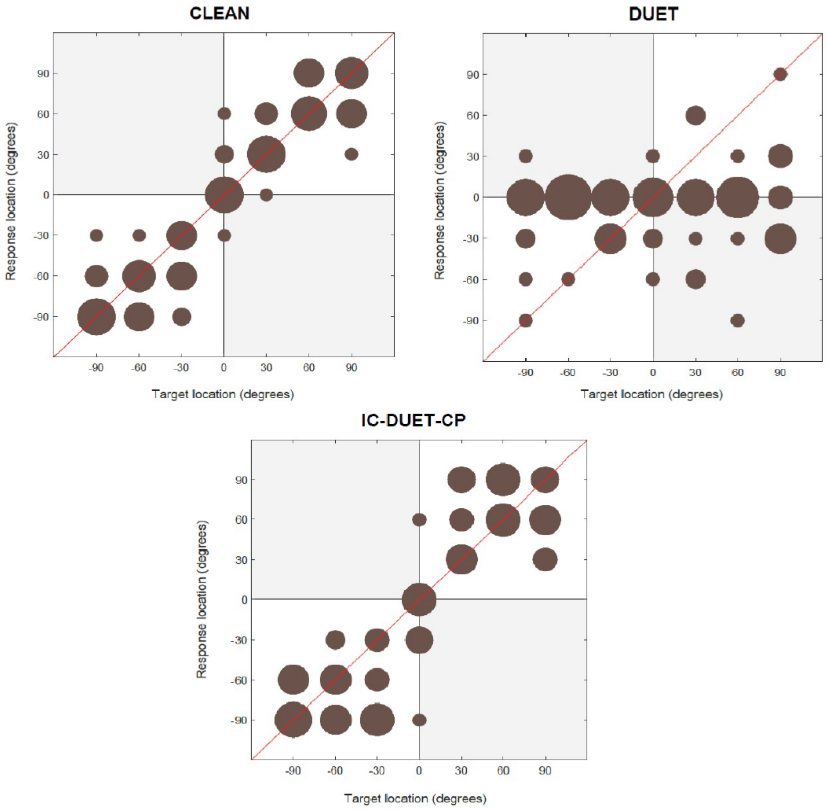

Figure 2: *Localization patterns for six listeners tested with the clean signal, DUET and IC-DUET-CP. The response location is plotted as a function of the target location. The area of each circle is proportional to the number of responses.*

of two presentations from each location. The listeners were instructed to identify the perceived virtual location of the male spoken sentence as accurately as possible by verbally indicating the number 1 to 7. A diagram showing the location of the subjects' head and the seven possible virtual locations of the male spoken sentence relative to their position was mounted on the wall directly in front of the subjects. The locations were numbered from 1–7 with -90^{o} corresponding to location 1 going clockwise in 30^{o} steps to $+90^{o}$ represented by location 7. Each subject produced a total of 14 responses (7 locations x 2 repetitions) per condition. To quantify the ability of the algorithm in retaining spatial cues in the signal estimates, we calculated the RMSLE between the azimuth of the presented stimulus location and the actual responses [21]:

$$\text{RMSLE}(k) = A\sqrt{\frac{1}{M}\sum_{i=1}^{M}(r_i - s_k)^2} \tag{17}$$

where A is the angular separation expressed in degrees, M is the number of responses, r_i is the listener's response to the i^{th} trial on which the source is presented, and s_k represents the source number.

3.5. Discussion

As shown in Table 1, the DUET and DUET-CP algorithms yield high values of SDR and SIR in the anechoic condition (ANE) tested. However, in reverberation (REV), there is a significant decrease in both criteria. Therefore, employing the dereverberation stage is essential before source separation. According to Table 1, when compared to the unprocessed reverberant mixtures, the DUET-CP increases SDR by 2.46 dB and SIR by 6.28 dB. Moreover, the IC-DUET-CP produces an SDR equal to 2.40, which is 2.96 dB greater than the SDR of the unprocessed mixtures and 0.5 dB greater than the SDR calculated after DUET was applied.

The proposed algorithm retains the quality and adds no distortion to the output signals. Furthermore, the IC-DUET-CP algorithm produced higher SIR scores than the scores corresponding to the unprocessed mixtures and the DUET outputs. The cue preservation stage also maintains the SDR and SIR values, while preserving the location of the sources. The pre-processing dereverberation stage increases the SRMR and segSRR by 1.36 dB and 2.30 dB respectively. Fig. 2 shows bubble plots for the six listeners who participated in the localization task. The most striking feature of the localization responses is the relatively high rate of azimuth confusions observed when the target source estimates were generated without the cue preservation post-processing stage. In this case, the observed RMS localization error averaged across all listeners was 63^{o}. In contrast, when the source estimates were extracted using cue preservation, most responses fell on the diagonal indicating correct sound source identification by the listeners. In this case, the observed RMS localization error was 32^{o}, while the error when the target speech signal was presented alone was equal to 23^{o}.

4. Conclusions

We proposed a blind source separation algorithm based on T-F masking that recovers the original sources in reverberation and preserves the spatial location of the sources. Preserving binaural cues increases the applicability of T-F binary masking strategies as it not only ensures improved speech intelligibility in reverberation but also enhances localization performance.

5. References

[1] A. S. Bregman, *Auditory Scene Analysis: The perceptual organization of sound*. Cambridge, MA : MIT Press, 1990.

[2] J. Blauert, *Spatial Hearing – The psychophysics of human sound localization*. Cambridge, MA : MIT Press, 1997.

[3] A. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acoustica*, vol. 86, no. 1, pp. 117–128, 2000.

[4] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Net.*, vol. 15, no. 5, pp. 1135–1150, 2004.

[5] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed., Kluwer Academic, Norwell, MA, pp. 181–197, 2005.

[6] D. L. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[7] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech perception of noise with binary gains," *J. Acoust. Soc. Am.*, vol. 124, no. 4, pp. 2303–2307, 2008.

[8] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking ," *IEEE Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[9] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation*, S. Makino, T. W. Lee, and H. Sawada, Eds. Dordrecht, The Netherlands: Springer, pp. 217–241, 2007.

[10] J. B. Allen, D. A. Berkley and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912–915, 1977.

[11] M. Jeub and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1732–1745, 2010.

[12] I. A. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 6, pp. 709–716, 2003.

[13] C. Faller and J. Merimma, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089, 2004.

[14] E. Vincent, R. Gribonval and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[15] T. H. Falk,C. Zheng, and W. YĊhan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[16] P. A. Naylor, N. D. Gaubitch and E. A. P. Habets, "Signal-based performance evaluation of dereverberation algorithms," *Journal of Electrical and Computer Engineering*, vol. 2010, pp. 1–5, 2010.

[17] A. Tsilfidis and J. Mourjopoulos, "Blind single-channel suppression of late reverberation based on perceptual reverberation modeling," *J. Acoust. Soc. Am.*, vol. 129, no. 3, pp. 1439–1451, 2011.

[18] A. Westermann, J. M. Buchholz and T. Dau, "Binaural dereverberation based on interaural coherence histograms," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 2767–2777, 2013.

[19] IEEE Subcommittee, "IEEE recommended practice speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.

[20] H. Kayser, S. D. Ewert, J. Anemuller, T. Rohdenburg, V. Hohmann and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP J. Appl. Signal Process.*, pp. 1–10, 2009.

[21] B. Rakerd and W. M. Hartmann, "Localization of sound in rooms. III: Onset and duration effects," *J. Acoust. Soc. Am.*, vol. 80, no. 6, pp. 1695–1706, 1986.