AN EVALUATION OF STEREO SPEECH ENHANCEMENT METHODS FOR DIFFERENT AUDIO-VISUAL SCENARIOS

Alexandra Craciun[1], *Christian Uhle*[2] *and Tom Bäckström*[1,2]

[1] International Audio Laboratories Erlangen, Friedrich-Alexander University (FAU), Erlangen, Germany
[2] Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany
Email: alexandra.craciun@audiolabs-erlangen.de

ABSTRACT

Following speech on TV or radio in the presence of interferers is sometimes challenging, in particular for the elderly and the hearing-impaired. To evaluate the performance of speech enhancement methods for such scenarios, we consider a stereo mixture composed of a speech signal and interfering sources. We apply different approaches to separate the mixture into two components, where the first component contains mainly speech (the desired signal) and the second component contains the rest of the mixture. An improved stereo signal is constructed by recombining these components such that speech gets emphasized with respect to the rest of the mixture and at the same time the amount of artifacts is kept to a minimum. Listening tests and objective measures show that the center extraction approach is in general the most effective, although it is sensitive to speaker positioning.

Index Terms— speech enhancement, center extraction, noise suppression, direct-ambient decomposition.

1. INTRODUCTION

Understanding speech in typically problematic scenarios when the recordings are corrupted by relatively loud interferers represents a common problem in TV and radio broadcasts. Particularly for the elderly, the hearing-impaired and the non-native speakers this represents a big inconvenience. The current paper examines different types of algorithms for speech enhancement, with the aim of adding insight on how well each algorithm improves the listening experience in such difficult cases.

The objective of this paper is to investigate the suitability of different classes of methods for the scenario described above. This is particularly important since by being developed as different strategies, it is not clear how the methods compare in quality and how sensitive they are to differences between scenarios. For the purpose of this evaluation we narrow down typical problematic scenarios to three conditions: i. speech recorded in reverberant environments, ii. speech recorded in noisy environments and iii. speech recorded in the presence of interferers panned off-center (different types of interferers panned to either side of the sound scene). In the following we consider that the input stereo mixture is composed of speech (the desired signal) and background (the rest of the mixture), where the background consists of all possible types of interferers, i.e., musical instruments, noise, other non-speech sounds or a mixture thereof.

For reverberant speech recordings, there already exist a few algorithms that estimate the direct and ambient (diffuse) signals, which are then used to enhance loudspeaker outputs [1], [2], [3]. For our investigation we chose [2], where the diffuse sound is obtained by least-squares estimation. This allows to generate statistically independent diffuse signals which correspond to independent loudspeaker signals.

For enhancement of noisy speech recordings, the typical problematic scenarios for audio-visual media are those where the signal-to-noise ratio (SNR) is relatively low or where the present speech components are rather weak. We therefore chose the improved minima controlled recursive averaging (IMCRA) [4] as noise estimation method due to its robustness under such adverse conditions.

For speech recorded in the presence of interferers from the side, we assume that speech is panned only to the center, while the interferers are panned predominantly off-center. For such a scenario, the aim is to attenuate the off-center components of the signal. In [5], [6] this is achieved by employing spectral weights based on power ratio functions, while in [7] the spectral weights are computed based on similarities between the channels of the stereo signal. For our tests we chose [7] due to its implementation simplicity and its efficient estimation of center and off-center components.

In this paper we chose separate algorithms to extract speech in three scenarios we identified as problematic. Given an input stereo mixture, we extracted the desired speech, as well as the background signal using the proposed algorithms. Following this, we computed an output signal as an additive mixture of the speech and the attenuated background signal. The attenuation was manually controlled such that the number of artifacts was kept to a minimum. The performance of the proposed algorithms was evaluated both subjectively, by means of listening tests and objectively, with commonly used perceptually-oriented measures. The results showed that the center extraction method in general outperforms the other methods. However, the method is sensitive to speaker positioning and thus leads to less improvement in case the speech is panned off-center.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

2. SIGNAL MODEL

We consider a stereo input signal $\mathbf{y}[n]$ containing speech corrupted by different types of background interferers:

$$\mathbf{y}[n] = \mathbf{x}[n] + \mathbf{b}[n], \tag{1}$$

where $\mathbf{y}[n] = [y_1[n] \;\; y_2[n]]^T$. Here $\mathbf{x}[n] = [x_1[n] \;\; x_2[n]]^T$ and $\mathbf{b}[n] = [b_1[n] \;\; b_2[n]]^T$ represent the stereo speech signal and the stereo background signal, respectively.

Our aim is to create an enhanced stereo mixture $\mathbf{y}_E[n]$ in which the interferer is attenuated by a factor c, where $c \in (0, 1)$. Since the true speech and background components are not available, we scale the corresponding estimated signals:

$$\mathbf{y}_E[n] = \widehat{\mathbf{x}}[n] + c \cdot \widehat{\mathbf{b}}[n], \tag{2}$$

where $\widehat{\mathbf{x}}$ is the estimated speech signal and $\widehat{\mathbf{b}}$ is the estimated background signal.

The estimation of the speech and background components is done in the time-frequency domain by applying a short-time Fourier transform (STFT) to the input mixture $\mathbf{y}[n]$ such that Eq. (1) becomes:

$$\mathbf{Y}(m, k) = \mathbf{X}(m, k) + \mathbf{B}(m, k), \tag{3}$$

where m and k are, respectively, the frequency and time indices. An estimate of the speech signal $\widehat{\mathbf{X}}$ can then be obtained by applying spectral weights $\mathbf{G}_{method}$ to the input mixture $\mathbf{Y}$:

$$\widehat{\mathbf{X}}(m, k) = \mathbf{G}_{method,X}(m, k) \cdot \mathbf{Y}(m, k). \tag{4}$$

The spectral weights $\mathbf{G}_{method,X}$ are computed by processing the input mixture $\mathbf{Y}$ with a speech enhancement method. Each method aims at suppressing a specific case of interferer, while retaining as much as possible of the speech components.

For the noise suppression and center extraction methods, the two estimated signals are obtained by magnitude modifications of the input mixture's spectra. Therefore, $\widehat{\mathbf{B}}$ is easily computed as the difference between mixture $\mathbf{Y}$ and the estimated speech signal $\widehat{\mathbf{X}}$. For the direct-ambient decomposition method, the estimated speech and background signals are obtained by combining the two channels of the input mixture. That is, the two estimated components involve not only magnitude modifications of the mixture $\mathbf{Y}(m, k)$, but also phase modifications thereof. The estimated background $\widehat{\mathbf{B}}$ is in this case obtained by weighing the input mixture $\mathbf{Y}(m, k)$ with the spectral weights $\mathbf{G}_{method,B}$ such that $\widehat{\mathbf{B}}(m, k) = \mathbf{G}_{method,B}(m, k) \cdot \mathbf{Y}(m, k)$.

The estimated speech and background components $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{B}}$, respectively, are then transformed back to time domain by using an inverse short-time Fourier transform (ISTFT) and recombined with a factor c (see Eq. (2)), which results in the desired enhanced mixture $\mathbf{y}_E[n]$.

3. SPEECH ENHANCEMENT ALGORITHMS

In this paper we aim to evaluate and compare methods which allow us to separate an input stereo mixture into a speech component and a background component containing the unwanted interferer(s). We mainly distinguish between three different classes of methods: center extraction, noise suppression and direct-ambient decomposition, which are briefly described in the following.

Center extraction methods aim to extract the signal at the center of the sound scene. Television talks often have speech played from the center of a stereo frame if the TV has a 3-channel frontal loudspeaker array [8]. When the camera captures a scene with more talkers, one can easily map the speech more to the left/right according to the position of the speakers in the scene. However, if the camera switches to only one talker positioned in the center of the screen, it would feel unnatural to the viewer if the speech would still come from the left/right because he would expect the audio cues to match the visual ones. That is why such speech signals are typically mapped to the center channel.

The noise suppression methods aim at constructing a signal where noise is attenuated. This typically fits a scenario where speech is difficult to understand due to a high level of noise in the recording. The speech signal can be extracted from the initial mixture by first estimating the noise. Then, based on this estimate, spectral weights for filtering the noisy components out of the initial mixture are derived. Thus the final output is a speech signal with less noisy interference.

The direct-ambient decomposition methods aim at attenuating the diffuse components of a mixture, while the direct ones remain unchanged. Diffuse sound is typically composed of a very large number of sound reflections in an enclosed space, which have equal spatial distribution and equal intensity at any location in this space [9]. By removing/attenuating the diffuse part of the mixture, we obtain a signal which is less distorted by the room impulse response.

3.1. Center Extraction

The center extraction algorithm extracts the signal components panned to the center of a sound scene by exploiting the magnitude similarities between the left and right spectra of the stereo recording of the sound scene [7]. The speech estimate is thus obtained by multiplying the input mixture by the spectral weights $\mathbf{G}_{CE}$, which are derived from the spectra of the side signal $S(m, k)$. $S(m, k)$ is computed as the absolute value of the difference between the left and right channel of the mixture:

$$S(m, k) = |Y_1(m, k) - Y_2(m, k)|. \tag{5}$$

To obtain a center signal with a certain degree of attenuation of the side components, we subtract a portion of the side channel from each channel of the mixture:

$$Y_{C,i}(m, k) = |Y_i(m, k)| - w|S(m, k)|, \tag{6}$$

where $i \in \{1, 2\}$ is the channel index of the input signal and w is a weighting factor which indicates how much of the side signal is subtracted. In the following, time and frequency indices are discarded when possible for brevity.

Using $Y_{C,i}$, we can construct a spectral energy weighting function to extract the center components:

$$G'_i = \begin{cases} \left(\frac{Y_{C,i}}{|Y_i|}\right)^2, & Y_{C,i} \geq 0 \\ 0, & Y_{C,i} < 0. \end{cases} \quad (7)$$

In addition, we limit the amount of attenuation between g_{min} and g_{max} in order to avoid too large amplifications or attenuations:

$$G_{CE,X,i} = \begin{cases} G'_i, & g_{min} < G'_i < g_{max} \\ g_{max}, & G'_i > g_{max} \\ g_{min}, & G'_i < g_{min} \\ 0, & G'_i = g_{min}. \end{cases} \quad (8)$$

The final speech and background estimates can then be computed as $\widehat{\mathbf{X}} = \mathbf{G}_{\text{CE,X}} Y$ and $\widehat{\mathbf{B}} = \mathbf{Y} - \widehat{\mathbf{X}}$, respectively.

3.2. Noise Reduction

The noise reduction method we selected to use in this paper is based on the improved minima controlled recursive averaging (IMCRA) [10], [11]. It was chosen due to its efficiency in typically difficult noise estimation conditions such as input signals with low signal-to-noise ratio (SNR), weak speech components or nonstationary noise environments.

The IMCRA noise estimation method is an enhanced version of Cohen's previously proposed MCRA method [4]. The modifications include minimum tracking for active speech, a bias compensation factor and a better speech presence probability estimator. In IMCRA, the noise estimates are obtained by averaging over previous spectral power values of the measured noisy signals. This is done by means of a time-varying frequency-dependent variable which gets updated according to the speech presence probability. The smoothed spectral power values are then weighted by a constant factor which compensates for the bias when speech is absent.

The noise signal power spectrum $\overline{\lambda}_d$ is smoothed by:

$$\overline{\lambda}_d(m,k+1) = \widetilde{\alpha}_d(m,k)\overline{\lambda}_d(m,k) + [1-\widetilde{\alpha}_d(m,k)]|Y(m,k)|^2, \quad (9)$$

where the term $\widetilde{\alpha}_d$ denotes the time-varying frequency-dependent smoothing factor and can be computed as:

$$\widetilde{\alpha}_d(m,k) = \alpha_d(m,k) + (1-\alpha_d(m,k))p(m,k). \quad (10)$$

Here α_d is a smoothing constant between 0 and 1, while $p(m,k) = P(H_1(m,k)|\gamma(m,k))$ is the conditional speech presence probability. $H_1(m,k)$ represents the hypothesis that speech is present at frequency bin m and time frame index k, while $\gamma(m,k)$ is the *a posteriori* SNR [10].

After calculating the estimated noise signal spectrum $\overline{\lambda}_d$, we derive the spectral weights G_{NR} for extracting the speech components by means of spectral subtraction. The spectral subtraction approach we use is based on the multiband spectral subtraction in [11], which we apply, in contrast to the original publication, over all frequencies in a time frame. As a first step, the segmental SNR_{seg} is computed as:

$$\text{SNR}_{seg} = \frac{||\mathbf{Y}||_2}{||\sqrt{\overline{\lambda}_d}||_2}, \quad (11)$$

which is required for determining the oversubtraction parameter α (see [11]). We can then calculate the noise subtraction gains for recovering the speech components as:

$$G_{NR,X} = \left(\frac{|\mathbf{Y}|^2 - \alpha|\overline{\lambda_d}|}{|\mathbf{Y}|^2}\right)^2. \quad (12)$$

3.3. Direct-Ambient Decomposition

For the direct-ambient decomposition, we chose the spatial decomposition method proposed by Faller [2]. The method is perceptually motivated and involves separating the direct and diffuse sound components by least-squares estimation. The advantage of the approach is that it results in statistically independent diffuse signals, which can be used to generate independent loudspeaker signals.

The input stereo signal is modeled as follows:

$$\begin{aligned} Y_1(m,k) &= S(m,k) + N_1(m,k) \\ Y_2(m,k) &= A(m,k)S(m,k) + N_2(m,k), \end{aligned} \quad (13)$$

where S represents the STFT of the direct sound component mapped to a certain direction by factor A and N_1 and N_2 correspond to the STFTs of the independent lateral reflection components. In [2] it is assumed that the lateral reflection components have equal power ($P_N = P_{N_1} = P_{N_2}$).

Based on Eq. 13 and the previous model assumptions, we construct a system of equations for P_S, P_N (short-time power estimates of S and N) and A, whose solution is used to compute the required weighting factors for the direct and diffuse components:

$$\begin{cases} P_{Y_1} = P_S + P_N \\ P_{Y_2} = A^2 P_S + P_N \\ \Phi = \frac{aP_S}{\sqrt{P_{Y_1} P_{Y_2}}}. \end{cases} \quad (14)$$

Here Φ is the normalized cross-correlation between the left and the right input channel. The weighting factors $w_1 \ldots w_6$ are then computed together with post-scaling factors c_S, c_{N_1} and c_{N_2} [2]. These post-scaling factors ensure that the powers of the estimated direct and diffuse components are equal to P_s and P_N, respectively. The estimated direct and ambient components can then be obtained as:

$$\begin{aligned} \widehat{S} &= c_S(w_1 Y_1 + w_2 Y_2) \\ \widehat{N}_1 &= c_{N_1}(w_3 Y_1 + w_4 Y_2) \\ \widehat{N}_2 &= c_{N_2}(w_5 Y_1 + w_6 Y_2), \end{aligned} \quad (15)$$

where m and k are omitted for sake of simplicity. The final speech and background estimates become $\widehat{\mathbf{X}} = [\widehat{S}\ A\widehat{S}]$ and $\widehat{\mathbf{B}} = [\widehat{N}_1\ \widehat{N}_2]$ respectively. Note that since the left and

		SAR			SDR			SIR			PESQ		
C (dB)		6dB	9dB	12dB	6dB	9dB	12dB	6dB	9dB	12dB	6dB	9dB	12dB
CE	0^o	16.0	13.1	11.4	15.6	12.7	11.0	27.7	24.4	22.6	4.3	4.0	3.7
	-30^o	15.7	11.7	8.8	15.2	11.1	8.1	26.9	23.2	20.7	4.3	3.2	3.4
NR	0^o	15.7	12.9	11.3	15.2	12.2	10.5	24.5	20.8	18.7	4.3	4.0	3.7
	-30^o	16.3	13.0	10.7	16.2	12.8	10.6	33.5	30.2	28.1	3.6	4.2	3.8
DAD	0^o	10.8	7.9	5.8	10.6	7.7	5.6	23.7	22.1	20.7	4.1	3.5	3.2
	-30^o	12.8	10.5	8.8	11.1	8.4	6.6	16.7	13.3	11.6	3.5	3.2	2.8

Table 1. Performance of the center extraction (CE), noise reduction (NR) and direct-ambient decomposition (DAD) methods in terms of Signal-to-Artifacts Ratio (SAR), Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR) and Perceptual Evaluation of Speech Quality (PESQ) grade.

right spectra of the mixture are scaled and added together (see Eq. 15), this method results in phase modifications of the input mixture.

4. PERFORMANCE EVALUATION

4.1. Test data

For the evaluation of the proposed speech enhancement methods, we created 8 different mixtures of speech and background (BG) signals. As speech we used the female and male German speech signals from the *EBU SQAM CD* [12], which were then panned to the center (0^o) and to the left (-30^o), resulting in 4 different speech signals. For the background we used 2 excerpts of cheering crowd and dense applause from the *Series 6000 General Sound Effects Library* [13]. The choice of the background signals was motivated by typical problematic scenarios in audio-visual media. We thus obtained a total of 8 mixtures of speech and background signals, all sampled at 48kHz.

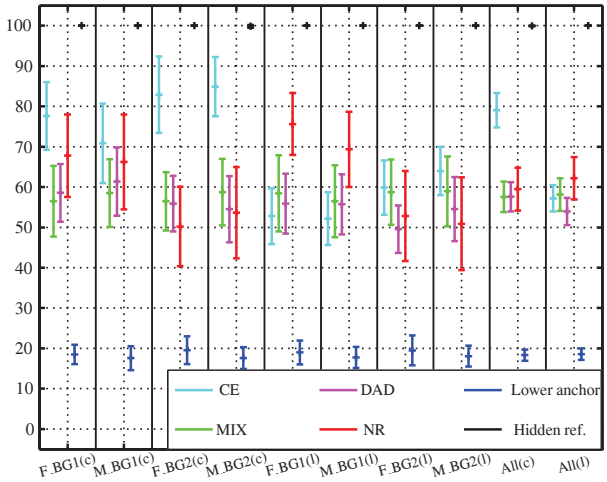

Fig. 1. MUSHRA results for center extraction (CE), noise reduction (NR), direct-ambient decomposition (DAD) and the original mixture (no BG attenuation) (MIX) for 8 mixtures (F: female German, M: male German, BG1: cheering crowd, BG2: dense applause) and 2 different speech pannings (c: center, l: left). The last 2 columns show the averaged results over all items panned to the center and to the left, respectively.

4.2. Objective Measures

For the objective evaluation we calculated the Signal-to-Artifacts Ratio (SAR), the Signal-to-Distortion Ratio (SDR) and the Signal-to-Interference Ratio (SIR) as defined in [14], as well as the Perceptual Evaluation of Speech Quality (PESQ) grade [15]. We tested the proposed speech enhancement methods for different attenuation levels C of the background: 6dB, 9dB and 12dB. Here $C = -10\log_{10}c$, where c is the background attenuation factor used in Eq. 2. The reference signal used for the comparison was a mixture of perfectly estimated speech and background signals, where the background was attenuated by the same values listed above. The results of the tests are shown in Table 1. We notice that the SAR, SDR and SIR decrease with increasing attenuation. This is often true also for PESQ, except for some cases where the source is panned to the left.

The reason for the decrease of the SAR, SDR and SIR measures for higher background attenuation is explained in the following. In general, none of the speech enhancement methods results in a perfect estimation of the speech or background signals. That is, an estimated source may contain components corresponding to the other source or may miss some segments, which were assigned to the incorrect source by the speech enhancement algorithm. In addition, distortions and artifacts can also appear. On one hand, by mixing the separated signals with no attenuation, the imperfections of the speech enhancement algorithms are not noticed because the two signals add to the exact original signal. On the other hand, when attenuating the background signal by a certain factor and adding it to the speech signal, the imperfections become more obvious. As a result, the audio quality of the enhanced mixture decreases the more the background signal is attenuated, which is consistent with the results in Table 1.

In general, the largest values for all measures are obtained for the center extraction (CE) and noise reduction methods (NR) for 6dB background attenuation. We note that CE performs best for a source at 0^o, while NR algorithms are better for a source at -30^o, which is to be expected since CE does not work so well for panned sources. Among all methods, the direct-ambient decomposition (DAD) is ranked last in performance. A particularly interesting behaviour is noticed when the speech source is panned to the left, where PESQ doesn't display a monotonically decreasing behaviour for CE and NR, but improves sometimes with increasing

attenuation. This suggests the necessity of performing subjective listening tests to better measure the quality of the proposed speech enhancement methods.

4.3. Listening Test

To evaluate the subjective audio quality of the enhanced mixtures, we carried out a MUSHRA listening test [16]. For the MUSHRA test, we chose to investigate closer only the case where the background was attenuated by 9dB. Our choice was motivated by the fact that for 6dB, the speech enhancement is less obvious, while for 12dB, the quality of the mixture degrades too much due to the fact that there is less masking for the artifacts resulted from the speech enhancement algorithm. A total of 20 listeners evaluated the proposed methods for the same 8 mixtures we analyzed in the objective evaluation. The test consisted of the following conditions: a lower anchor (3.5kHz low-pass version of the original mixture), a hidden reference (perfectly separated speech and BG with 9dB attenuation of the BG), the 3 mixtures created with our proposed methods and the original mixture (no BG attenuation). We asked the listeners to give an overall audio quality grade for each mixture taking into consideration the audio quality of both the speech and the BG with respect to the reference.

The means and the 95% confidence intervals can be seen in Figure 1. We notice that for the case when the source is panned to the center, CE was preferred, with ratings of good to excellent. However, for the case when the source is panned to the left, NR was graded better or almost as good as CE. Interestingly, the users graded the DAD often nearly as low as the unattenuated original mixture. This is due to the fact that for DAD the estimated speech signal contains a lot of BG components. Since only the BG is attenuated, this leads to unpleasant distortions and artifacts in the resulting mixture, which were perceived as less pleasant by the listeners.

5. CONCLUSIONS

In this paper we evaluated different speech enhancement methods for stereo signals which contain a mixture of speech and undesired interferers. The proposed methods are based on the decomposition of the stereo input into a speech and a background component, followed by the remixing of the speech with an attenuated version of the background. We investigated three different methods for the decomposition and three different levels of background attenuation.

According to the results of both objective and subjective tests, we can conclude that the center extraction and the noise reduction methods perform quite well under the right assumption (source panned to center and predominantly noisy background, respectively). The direct-ambient decomposition was found to be less suitable since it was graded roughly as low as the mixture with no background attenuation. In addition, the SAR, SDR, SIR and PESQ objective measures showed that in general the quality of the mixture deteriorates as the background attenuation becomes larger, in particular when the attenuation is larger than 9dB.

REFERENCES

[1] V. Pulkki, "Directional audio coding in spatial sound reproduction and stereo upmixing," *AES 28th International Conference*, June 2006.

[2] C. Faller, "Multiple-loudspeaker playback of stereo signals," *Journal of the Audio Engineering Society*, vol. 54, no. 11, pp. 1051–1064, 2006.

[3] C. Avendano and J.-M. Jot, "A frequency-domain approach to multichannel upmix," *J. Audio Eng. Soc*, vol. 52, no. 7/8, pp. 740–749, 2004.

[4] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, 2002.

[5] C. Uhle and E. Habets, "Subband center scaling using power ratios," in *AES 53rd International Conference*, London, UK, January 2014.

[6] C. Uhle, "Center signal scaling using signal-to-downmix ratios," in *DAFx 2013*, Maynooth, Ireland, September 2013.

[7] C. Uhle, S. Finauer, P. Gampp, O. Hellmuth, P. Prokein, and C. Stoeklmeier, "Method and apparatus for decomposing a stereo recording using frequency-domain processing employing a spectral weights generator," May 2014, Patent US20140119545 A1.

[8] S. R. Alten, *Audio in Media (10th Edition)*, Cengage Learning, 2013.

[9] J. Blauert and N. Xiang, *Acoustics for Engineers: Troy Lectures (2nd Edition)*, Springer, 2009.

[10] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[11] P. C. Loizou, *Speech Enhancement, Theory and Practice*, Taylor & Francis Group, 2007.

[12] EBU SQAM-CD, "Sound quality assessment material recordings for subjective tests," 2008.

[13] Series 6000 General Sound Effects Library, "CD #6013," 1992.

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

[15] ITU-R Recommendation P.862, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assess of narrow-band telephone networks and speech codecs," *International Telecommunications Union*, February 2001.

[16] ITU-R Recommendation BS.1534-2, "Method for the subjective assessment of intermediate quality level of audio systems (MUSHRA)," *International Telecommunications Union, Geneva, Switzerland*, June 2014.