Subjective and Objective Quality Assessment of Single-Channel Speech Separation Algorithms

P. Mowlaee[†], R. Saeidi[⋆], M. G. Christensen[⋆†], and R. Martin[†]
[†] Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum (RUB), Bochum, Germany
[⋆] School of Computing, University of Eastern Finland, Joensuu, Finland
[⋆†] Dept. of Architecture Design and Media Technology, Aalborg University, Aalborg, Denmark
{pejman.mowlaee,rainer.martin}@rub.de, rahim.saeidi@uef.fi, mgc@imi.aau.dk

***Abstract*—Previous studies on performance evaluation of single-channel speech separation (SCSS) algorithms mostly focused on automatic speech recognition (ASR) accuracy as their performance measure. Assessing the separated signals by different metrics other than this has the benefit that the results are expected to carry on to other applications beyond ASR. In this paper, in addition to conventional speech quality metrics (PESQ and SNR_{loss}), we also evaluate the separation systems output using different source separation metrics: blind source separation evaluation (BSS EVAL) and perceptual evaluation methods for audio source separation (PEASS) measures. In our experiments, we apply these measures on the separated signals obtained by two well-known systems in the SCSS challenge to assess the objective and subjective quality of their output signals. Comparing subjective and objective measurements shows that PESQ and PEASS quality metrics predict well the subjective quality of separated signals obtained by the separation systems. From the results it is observed that the short-time objective intelligibility (STOI) measure predict the speech intelligibility results.**

***Index Terms*—Single-channel speech separation, subjective and objective quality assessment.**

I. Introduction

In many speech applications, speech quality evaluation is a crucial step in the development of new algorithms. This is particularly important in the field of blind source separation (BSS), speech enhancement and signal processing for hearing aids. In recent years, creating meaningful performance assessments of existing speech enhancement or speech separation algorithms has been considered as an important problem. The reasons for this difficulty is mainly due to the following: lack of an appropriate distortion measure to perform a reliable objective quality assessment between different algorithms, differences between the testing methodologies used by researchers in the field, and lack of an unambiguous exact definition to describe what really indicates good speech quality in a certain application [1]. As a result, it still remains unclear how one can compare the output of different single-channel speech separation (SCSS) systems in a fair manner. To this end, we conduct multiple objective and subjective evaluations on the output of two SCSS systems to assess their performance quality from different perspectives. In SCSS application, given a mixture of speech signals, we are interested in recovering all sources.

Attempts have been made to find reliable and efficient performance metrics to evaluate separation and enhancement algorithms [2], [3]. For example, previous studies in SCSS reported the separation performance in terms of word error rate (WER) [4] or signal-to-noise ratio (SNR) [5]. Evaluating the separation performance using automatic speech recognition (ASR) systems depends on several key parameters other than the separation system itself, namely: extracted features, language models, and acoustic models [6]. Similarly, relying only on SNR results obtained by a separation method can be illusive [7], [8]. Modified versions of common SNR measure were suggested in [9] for the evaluation of blind audio source separation algorithms. In particular, it was suggested to use three different measures: signal-to-interference ratio (SIR), signal-to-distortion ratio (SDR), and signal-to-artifact ratio (SAR) each reflecting one aspect of the resulting distortion attributed by their selected separation algorithm. It was concluded that the new measures better predicted separation performance compared to SNR [9].

Although different measures have been utilized to report the overall performance of SCSS system (see e.g. [10] for a detailed list of measures used in the literature), still none of them can fully characterize the human perception of quality and intelligibility which determines what we consider good results. In this paper, we base our investigation on subjective evaluations conducted on the separated signals obtained by the two systems [11], [12] that participated in the single-channel speech separation and recognition challenge [4]. We look for objective measures resemble the subjective evaluation. To the best of our knowledge, this is the first attempt at reporting different quality and intelligibility metrics together with ASR accuracy for evaluating the SCSS performance. This enables us to have a comprehensive comparison between the separation results obtained by existing separation methods already presented in the literature.

II. Subjective and Objective Measures

The distortion types produced by a typical speech separation algorithm are often classified into three classes: speech distortion, cross-talk interference, and artifacts (not correlated to any source) [9]. Using a single criterion cannot describe these different types of distortions. Different measures have been employed for predicting the quality of the separated signals. In [5], the authors suggested to combine computational auditory scene analysis with the ITU-T P.563 algorithm as their objective quality assessment of speech. They showed that the proposed approach achieves substantially good subjective perceived speech quality of separated speech. Perceptual evaluation of speech quality (PESQ) [13] was employed to assess the separation quality of single-channel [14] and multi-channel [15] methods.

In [14], the authors conducted subjective listening test using MUlti-Stimulus test with Hidden Reference and Anchors (MUSHRA) test [16] to evaluate the separation performance of different SCSS methods in terms of their perceived speech quality. Such a test enables the subjects to carry out simultaneous comparison between the methods directly. Similarly, in [17], a MUSHRA-based subjective test protocol was proposed for predicting the subjective scores obtained by different audio source separation algorithms.

The work of Pejman Mowlaee was funded by the European Commission within the Marie Curie ITN AUDIS, grant PITNGA-2008-214699. The work of Rahim Saeidi was supported in part by a scholarship from NOKIA foundation.

In [1], speech intelligibility was defined as the probability of correct recognition when context is available. Human speech recognition (HSR) is a two-step procedure [1]: (1) decoding the syllable as independent phone units extracted from the acoustic speech signal, and (2) correcting and fill the missing information via using context [1]. For predicting the resulting speech intelligibility performance, the speech intelligibility index (SII) is often used as a measure [18]. It calculates the weighted SNR in frequency domain based on a critical-band filtered signal representation. Finally, a standard for measuring the intelligibility for noisy speech mixtures was developed in [19].

The aforementioned measures would take into account different principles in assessing the performance of the enhanced signals obtained from a speech separation algorithm. As an example, results in [7] showed that the conventional SNR-based measures are illusive for evaluating different methods because they correlate poorly with subjective assessments. They are very sensitive to experimental conditions and result in artifacts due to fractional delays between the signals to evaluate [9]. Similarly, the modified SNR measures suggested in [9] largely depend on the number of delays and time frames chosen for the signal decomposition. We employ subjective and objective measures described in subsequent subsections to look at systems performance from multiple points of view.

A. Subjective Measures

1) MUSHRA: To assess the perceived speech quality of the separated output signals obtained by different separation methods, we consider a subjective listening test using the MUSHRA test as described in [16].

In the following, we present subjective and objective measures. These metrics have been introduced in diverse studies scattered in the literature but have never been reported together for assessment of single-channel speech separation algorithms.

2) Intelligibility test: Following the principle and the standard described in [19], here, we consider a speech intelligibility test to assess the resulting speech intelligibility of the separated signals obtained by different separation methods.

B. Objective Measures

1) STOI [3]: short-time objective intelligibility (STOI) measure was shown to have better correlation with speech intelligibility compared to other existing objective intelligibility models [3].

2) Cross-talk [20]: An ideal separation system would filter out any trace of the interfering speaker signal in the mixture.

3) PESQ [13]: PESQ is among the most widely used objective assessment tools in speech enhancement literature which correlates well with subjective listening scores [7].

4) SNR_{loss} [2]: This measure was found appropriate in predicting speech intelligibility in different noisy conditions by yielding a high correlation for predicting sentence recognition in noisy conditions ($r = -0.82$ higher than $r = 0.77$ for PESQ).

5) BSS EVAL metrics [9]: blind source separation evaluation (BSS EVAL) metrics have been quite standard in source separation. The metrics are:

- Signal-to-distortion ratio (SDR): measures the amount of distortion introduced by the output signal and is defined as the ratio between the energy of the clean signal, and that of the distortion.
- Signal-to-interference ratio (SIR): is defined as the ratio of the target signal power to that of the interference signal and measures the amount of undesired interference signal still remained in the separated signal.
- Signal-to-artifact ratio (SAR): measures the quality in terms of absence of artificial noise.

6) PEASS [17]: Perceptual evaluation methods for audio source separation (PEASS) were adopted for the signal separation evaluation campaign (SiSEC) evaluation suggested in [17]. They suggested four quality scores: overall perceptual score (OPS), target-related perceptual score (TPS), interference-related perceptual score (IPS) and artifacts-related perceptual score (APS). OPS measures how close is the separated signal, as it is, to the clean signal, TPS implies that how close is the separated signal to the clean one, IPS measures the interference cancellation in the separated signal, and finally APS shows how close is the enhanced signal to the clean one in terms of having no artifacts.

7) WER: The metric shows the ASR accuracy and was used as the only performance measure in the SCSS challenge [4].

III. Experimental Results

A. Dataset and Benchmark methods

For performance evaluation, we use a number of excerpts taken from the corpus in [4]. For performance evaluation, we use a number of excerpts taken from the corpus in [4]. In our performance evaluation, we assumed that the reference clean signal and the mixture are available for comparison purposes. The corpus in [4] is assumed to have no additional noise or reverberation. In our experiments, we used two methods which participated in the SCSS challenge: the "IBM super-human speech recognition system" proposed in [11] and the "speaker-adapted eigenvoice full system" proposed in [12] both working at a sampling rate of 16 kHz. The super-human speech recognition system [11] is based on factorial HMMs and was the top performing system in the SCSS challenge [4] having an average word recognition accuracy of 78.4%. The speaker-adapted eigenvoice system is also a model-based approach performing in the median range of 48% average accuracy among those reported by other participants in the challenge. In our experiments, we had access only to limited separated clips for the system in [11], where the authors in [12] supplied their separated signals on the whole GRID corpus. Therefore, we conducted our experiments on a limited set of clips which still covered different mixing scenarios: difference gender (DG), same gender (SG) and same talker (ST) at different signal-to-signal ratios (SSRs) satisfying the sake of generality[1]. For statistical analysis, pairwise *t-tests* are conducted to see if the subjective listening and objective results are statistically significant. The p-value reflects if one method achieves a statistically significant improved performance compared to another.

B. Subjective Measures

1) MUSHRA test: For the listening tests, the segments were selected from the available clips as representatives for the separated signals. Additionally, we chose two more segments as the hidden reference and anchor point. The hidden reference is used to ensure the consistency of subjects while performing the listening test. For hidden reference, we chose the clean reference signal. As our anchor point, we chose the speech mixture which reflects how difficult was to perceive speakers signals directly from their mixture.

Seven untrained listeners participated in the test (the authors were not included). The listeners were asked to rank the separated signals relative to a known reference on a scale of 0 to 100 by entering their results using computer graphical interface. Subjects had the possibility to listen to the audio segments as many times as they

[1]The clips are Clip 1: target sp6:bwba masker sp30:pgah6a (mixed at -3 dB), Clip 2: target sp14:lwax8s masker sp22:bgwf7n (mixed at 0 dB), Clip 3: target sp33:bwid1a masker sp33:lgii3s (mixed at -6 dB) and Clip 4: target sp5:swah6n masker sp5:bbir4p (mixed at 0 dB) signal-to-signal ratio Clip 5: target sp31: lwai5a masker sp31: pgin4p (mixed at 0 dB).

TABLE I

TWO SYSTEM COMPARISON WITH DIFFERENT METRICS ON FOUR CLIPS FROM GRID CORPUS. SYSTEMS ARE S1: HERSHEY [11] AND S2: WEISS [12]. METRICS ARE SHORT-TIME OBJECTIVE INTELLIGIBILITY (STOI) MEASURE [3], CROSS-TALK [20], PERCEPTUAL EVALUATION OF SPEECH QUALITY (PESQ) [13], SIGNAL-TO-INTERFERENCE RATIO (SIR) [9], SIGNAL-TO-ARTIFACT RATIO (SAR) [9], SIGNAL-TO-DISTORTION RATIO (SDR) [9], SNR_{LOSS} MEASURES [2], OVERALL PERCEPTUAL SCORE (OPS), TARGET-RELATED PERCEPTUAL SCORE (TPS), INTERFERENCE-RELATED PERCEPTUAL SCORE (IPS), ARTIFACTS-RELATED PERCEPTUAL SCORE (APS). SIR, SAR AND SDR ARE EXPRESSED IN DECIBELS. EACH CLIP IS CHARACTERIZED BY ITS MIXING SSR LEVEL AND THE MIXING SCENARIO: DIFFERENT GENDER (DG), SAME GENDER (SG) AND SAME TALKER (ST). STATISTICALLY SIGNIFICANT IMPROVED RESULTS OF SYSTEM S_1 WITH RESPECT TO S_2 ARE HIGHLIGHTED IN SHADED BOLDFACE FONT (EXCEPT FOR SDR MEASURE WHERE THE IMPROVED RESULTS OF S_2 ARE STATISTICALLY SIGNIFICANT WITH RESPECT TO S_1). THE NON-SIGNIFICANT ONES ARE HIGHLIGHTED WITH BOLDFACE FONT.

		Target								Masker								
	Criterion	Clip 1 (SG -3dB)		Clip 2 (DG 0dB)		Clip 3 (ST -6dB)		Clip 4 (ST 0dB)		Clip 1 (SG -3dB)		Clip 2 (DG 0dB)		Clip 3 (ST -6dB)		Clip 4 (ST 0dB)		p-value
		S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	
	STOI	**0.77**	**0.77**	0.80	**0.82**	**0.74**	0.51	**0.85**	0.48	**0.83**	**0.83**	**0.82**	0.70	**0.79**	0.43	**0.78**	0.49	**<0.05**
	Cross-talk	**11.50**	13.30	**10.30**	11.90	4.30	**2.30**	**4.40**	9.40	10.10	**5.90**	**10.30**	10.50	**13.80**	17.30	12.40	**10.10**	>0.05
	PESQ	**2.40**	0.70	**2.38**	1.48	**2.41**	1.73	**2.20**	1.20	**2.30**	1.70	**1.20**	1.00	**2.91**	0.89	**2.40**	1.70	**<0.05**
	SNR_{loss}	**0.96**	0.98	0.99	**0.89**	**0.92**	0.98	**0.91**	0.98	**0.93**	0.99	**0.92**	0.97	**0.89**	0.96	0.97	**0.95**	>0.05
BSS EVAL	SIR	0.13	**14.06**	8.52	**25.2**	1.31	**2.18**	4.49	**16.53**	10.84	**13.64**	0.69	**7.01**	**8.60**	6.12	**8.12**	-8.37	>0.05
BSS EVAL	SAR	**71.23**	-1.74	-1.70	**1.21**	0.98	**1.54**	-4.45	**-4.41**	**0.52**	-1.58	**-1.43**	-6.1	**0.23**	-7.59	**-3.04**	-6.94	>0.05
BSS EVAL	SDR	-9.16	**-5.79**	-8.15	**-3.05**	-8.64	**-6.04**	-7.01	**-3.66**	-10.32	**-4.87**	-8.56	**-8.19**	-8.67	**-3.41**	-10.59	**-4.47**	**<0.05**
PEASS	OPS	**43**	20	**51**	38	**59**	32	**50**	15	**69**	29	**48**	20	**73**	19	**65**	28	**<0.05**
PEASS	TPS	71	**79**	**70**	64	**62**	35	**77**	23	60	**62**	**68**	55	**63**	25	**57**	49	**<0.05**
PEASS	IPS	55	**65**	**81**	77	**82**	76	**79**	63	**85**	74	**76**	65	**86**	71	**85**	79	**<0.05**
PEASS	APS	**94.2**	15	52	**62**	**60**	24	**52**	9	**69**	21	**59**	13	**71**	9	**62**	16	**<0.05**

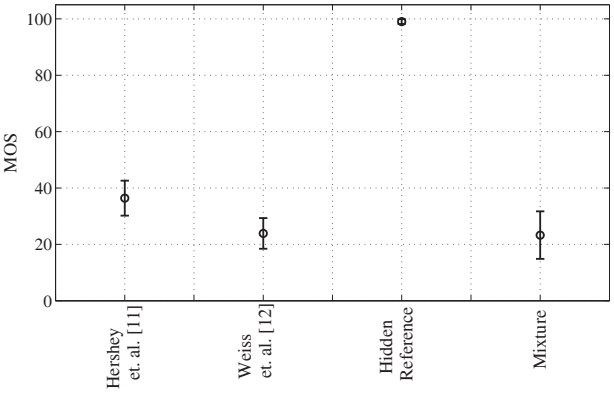

Fig. 1. Results of the MUSHRA listening test for different separation methods averaged over all excerpts and listeners. Error bars indicate 95% confidence intervals.

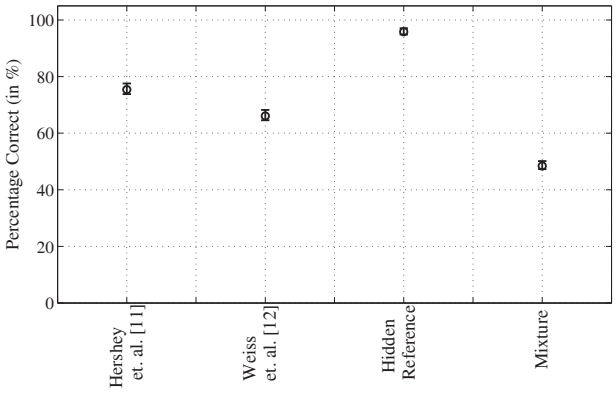

Fig. 2. Speech intelligibility test results. The calculated percentage of correct keywords is averaged over all excerpts and all listeners. Error bars indicate confidence intervals.

wanted. The segments were played in random to each subject.

Figure 1 shows the mean opinion score (MOS) results calculated and averaged over all clips and listeners. The mean and confidence intervals for each method is calculated following the standard as described in ITU-R BS.1534-1 [16]. It is observed that the method in [11] achieves a better MOS result compared to the one in [12]. We observe that the maximum and minimum scores were obtained at hidden reference and speech mixture, respectively, as expected. From the significance test results, we observed that the IBM separation method achieves statistically significant improvement compared to eigenvoice speaker-adapted method in [12].

2) Speech intelligibility: We followed the routine provided in [19] to conduct a speech intelligibility test. The same clips were used as described for MUSHRA test. The listeners were asked to identify color, alphabet letter, and digit number spoken during each of the played segments. On average, it took 15 minutes per listener to complete the test. The percentage of correct keywords were calculated and averaged over all clips and all listeners. Figure 2 shows the speech intelligibility performance in the form of the mean value together with a confidence interval. The intelligibility results are averaged over the listeners and the clips played for the listeners. As expected, the mixed signal and the hidden reference achieve the lowest and the highest speech intelligibility performance among the clips. We observe that the method in [11] achieves statistically significant improvement compared to the one in [12]. This is also clear from Fig. 2 where the confidence intervals of the two methods do not overlap.

3) Compared to mixture: To measure the effectiveness of the separation methods, it is useful to measure how much improvement is achieved after applying the separation method compared to that directly given by the mixture. From paired test results, it was observed that only [11] achieves the statistically significant improvement. This is visible from the MUSHRA test results illustrated in Fig. 1 as the confidence intervals of [12] overlaps with those of mixture. For speech intelligibility, both methods attain a statistically significant improvement compared to the mixture which is in line with intelligibility results shown in Fig. 2 as both methods have higher speech intelligibility compared to mixture level and their confidence intervals do not overlap with those of mixture.

C. Objective Measures

The separation results together with the significance level in the form of p-values, are summarized in Table I for the selected clips. The following observations are made.

From the PEASS quality metrics (OPS, TPS, IPS and APS) and PESQ results, it is observed that the IBM system achieves statistically significant better performance compared to [12]. We also conclude that these measures follow the MUSHRA results and predict the speech quality obtained by different methods. Comparing the separation systems performance to mixture case, it was observed that the aforementioned measures well predict MUSHRA measurements.

Speaker-adapted eigenvoice system often achieves higher scores in terms of SIR compared to IBM system. Such improved interference rejection performance obtained by [12] is achieved at the price of low SAR results, introducing more artifacts. This implies that a separation quality with less cross-talk is feasible when introducing more artifacts. Similar result on trade-off between improvement in interference rejection (SIR) versus achieving a lower amount of artifacts (SAR) were reported in [9]. Finally, according to the SDR results shown in Table I, only the system in [12] achieves statistically significant better performance in comparison to mixture. As the separated signals were not time-aligned nor scale-aligned with respect to the original signals, the SDR scores are negative. SDR has no preference over interference signal or noise power. Therefore, the same level of each will degrade the SDR metric by the same amount.

The difference in speech intelligibility performance between the two systems is significant as shown in Fig. 2. From the statistically significant STOI results, we conclude this metric well predicts the speech intelligibility scores. The SNR_{loss} results indicate insignificance difference in performance, therefore, we conclude that it is not a reliable predictor for predicting speech intelligibility obtained by SCSS algorithms.

The gap between the ASR results in [4] is larger than their gap in the speech intelligibility results presented in Fig. 2. There are three reasons for this; (1) HSR is based on utilizing contextual information like meaning conveyed by words or sentence [1]. Using such high-level knowledge enables listeners to compensate some missing information, accordingly results in higher accuracy compared to ASR, (2) ASR systems emphasis on word and language models as a method of increasing their recognition accuracy [1], and (3) ASR is based on template matching of spectral features. Possible degradation in one frequency affects the whole template. As a result, ASR systems are not robust to noise and reverberation. In contrast, human partially recognize speech by obtaining frequency-independent speech features, making HSR robust at low SNRs [1].

IV. Conclusion

Performance evaluation of different speech separation algorithms has proven to be a difficult task since current existing assessment tools do not fully reflect the quality of the resulting separated signals. In this paper, we employed different objective and subjective measures and evaluated the separation performance obtained by two well-known single-channel speech separation methods on a limited set of available clips. From the results it was observed that PESQ as objective measure correlates with the MUSHRA results. Based on the presented results, it was observed that PESQ and PEASS quality measures could well predict separation quality. Furthermore, the STOI measure seem to be well suited to predict the speech intelligibility result. These measure are, therefore recommended to evaluate single-channel speech separation algorithms. The results in terms of SIR and cross-talk measures indicated that there is no preference between the two methods in terms of reducing traces of the interfering signal in the separation output.

Acknowledgment

Authors would like to thank Dr. Ron Weiss and Prof. Dan Ellis for their assistance in sharing their data and Dr. Emmanuel Vincent for his helpful discussions concerning the BSS EVAL implementation.

References

[1] J.B. Allen, "How do humans process and recognize speech?," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 567 –577, Oct. 1994.

[2] J. Ma and P. C. Loizou, "SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech," *Speech Communication*, vol. 53, no. 3, pp. 340 – 354, 2011.

[3] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2125 –2136, Sept. 2011.

[4] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

[5] P. Li, Y. Guan, B. Xu, and W. Liu, "Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 2014 –2023, Nov. 2006.

[6] O. Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," *Speech Communication*, vol. 49, no. 5, pp. 336 – 347, 2007.

[7] H. Yi and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[8] L. Di Persia, M. Yanagida, H. L. Rufiner, and D. Milone, "Objective quality evaluation in blind source separation for speech recognition in a real room," *Signal Process.*, vol. 87, pp. 1951–1965, Aug. 2007.

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

[10] P. Mowlaee, *New Stategies for Single-channel Speech Separation*, Ph.D. thesis, Department of Electronic Systems, Aalborg University, Aalborg, Denmark, Dec. 2010.

[11] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, Jan 2010.

[12] R. J. Weiss and D. P. W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 16–29, 2010.

[13] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 749–752, Aug. 2001.

[14] P. Mowlaee, M. Christensen, and S. Jensen, "New results on single-channel speech separation using sinusoidal modeling," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 5, pp. 1265 – 1277, 2011.

[15] L. Di Persia, D. Milone, H. L. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Processing*, vol. 88, no. 10, pp. 2578 – 2583, 2008.

[16] "ITU-R BS.1534-1, Method for the subjective assessment of intermediate quality level of coding systems.," 2001.

[17] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2046 –2057, 2011.

[18] "ANSI S.3.5-1997, American national standard methods for the calculation of the speech intelligibility.," 1997.

[19] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Commun.*, vol. 49, pp. 402–417, 2007.

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