Subjective Performance Evaluation of Single-channel Speaker-conditioned Target Speaker Extraction Algorithms for Complex Acoustic Scenes

Ragini Sinha[*], *Ann-Christin Scherer*[*], *Simon Doclo*[*,†], *Christian Rollwage*[*], *Jan Rennies*[*]

[*] Fraunhofer Institute for Digital Media Technology IDMT, Oldenburg Branch for Hearing, Speech and Audio Technology HSA, Germany
[†] Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany
Email:ragini.sinha@idmt.fraunhofer.de

Abstract

This study investigates the performance of speaker-conditioned target speaker extraction algorithms. While previous studies mostly focused on instrumental measures, this paper employs three different subjective performance measurement methods for two algorithms, namely: paired comparison, speech intelligibility measurement, and categorically scaled listening effort. The subjective evaluations with 15 normal-hearing subjects for different mixtures show a clear benefit of the time-domain-based algorithm compared to the magnitude-based algorithm and the unprocessed mixtures, i.e., it is clearly preferred in direct comparisons and produces significantly lower listening effort and better intelligibility. The time-domain-based algorithm also improves SRTs compared to the unprocessed mixtures even though unprocessed reference SRTs were very low. In contrast, the magnitude-based algorithm shows no improvement over the unprocessed mixtures in any evaluation method.

1 Introduction

The goal of speaker extraction is to extract a target speaker from a complex acoustic mixture consisting of multiple sounds, such as the mixture of overlapping speech from multiple speakers and background noise [1]. This is highly relevant for several real-world applications including teleconferencing, voice-enabled devices, and hearing aid devices. In principle, target speaker extraction can be achieved by utilizing blind source separation [2–4] to extract all sources from the mixture and select the source corresponding to the target speaker. Alternatively, speaker-conditioned target speaker extraction algorithms [5–12] directly estimate the target speaker from the mixture by exploiting auxiliary information about the target speaker. Different types of auxiliary information were proposed, such as reference speech [5–10], visual information [11], and spatial information [12].

Several single-channel target speaker extraction algorithms [5–10] have been proposed and have shown impressive results when their performance is assessed in terms of speech quality and intelligibility [13–16] using various instrumental measures. Although these measures are commonly used due to their fast computation and low cost, none of them can fully reflect the human perception of speech quality and intelligibility, which is why formal listening tests remain the gold standard for algorithm evaluations. To the best of our knowledge, no previous study has evaluated the performance of target speaker extraction employing different methods of formal listening tests. Thus, it is unknown which experimental method is best suited for this task, especially since different methods differ with respect to the applicable range of signal-to-noise ratios (SNRs).

This study focuses on the subjective performance evaluation of speaker-conditioned target speaker extraction algorithms for normal-hearing listeners employing 3 different methods, i.e., paired comparisons of differently processed stimuli at a fixed SNR of 0 dB [17], adaptive measurements of speech recognition thresholds (SRTs), i.e., SNRs corresponding to 50% speech intelligibility which typically result in very low (negative) SNRs [18], and categorically scaled perceived listening effort, which can be measured across a broad range of SNRs [19]. In this study, we consider 2 different single-channel speaker-conditioned target speaker extraction algorithms. The potential for enhancing the perception of the target speaker's voice was measured for both considered algorithms over a broad range of SNRs using three experimental methods for different acoustic conditions consisting of either one or two interfering speakers, with or without sex differences to the target speaker.

2 Target Speaker Extraction

A speaker-conditioned target speaker extraction algorithm consists of an embedder network and a separator network. The embedder network generates an embedding from the reference speech of the target speaker, which is used to guide the separator network to discriminate between the target speaker and interfering speakers present in the mixture. The embedder and separator networks can be optimized either separately [5] or jointly [6–9] to perform the speaker extraction in the spectral domain [5] or in the time-domain [7, 9].

Consider a time-domain signal $y(n)$ consisting of a mixture of the target speaker $x_t(n)$ and I interfering speakers, i.e.,

$$y(n) = x_t(n) + \sum_{i=1}^{I} x_i(n), \tag{1}$$

where $x_i(n)$ denotes the speech signal of the i-th interfering speaker and n denotes the discrete-time index. The goal is to extract the target speaker by exploiting the reference speech $r_t(n)$. This is achieved by first computing a speaker embedding from $r_t(n)$ using the embedder network, and then utilizing it to guide the separator network towards estimating the target speaker from the mixture, i.e.,

$$\hat{x}_j(n) = \boldsymbol{\phi}^{sep}(y(n), \boldsymbol{\phi}^{emb}(r_t(n))), \tag{2}$$

The Oldenburg Branch for Hearing, Speech and Audio Technology HSA is funded in the program »Vorab« by the Lower Saxony Ministry of Science and Culture (MWK) and the Volkswagen Foundation for its further development. This study was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Projektnummer 352015383 - SFB 1330 A1 and B2.

where ϕ^{emb} and ϕ^{sep} denote the functional model of the embedder network and the separator network, respectively. Among the considered 2 target speaker extraction algorithms, the first algorithm is inspired by [5] (magnitude-based) which performs the extraction in the spectral domain by computing the real-valued mask and separately optimizes the embedder and the separator network. Whereas the second algorithm [9] (time-domain-based) performs the speaker extraction in the time domain and jointly optimizes the embedder and separator networks. Both algorithms were trained for mixtures of 2 speakers, mixtures of 3 speakers, and noisy mixtures of 2 speakers together on the same dataset generated using the WSJ0 [20] and the WHAM [21] corpora as in [9]. Based on instrumental evaluations (SI-SDR (in dB) [15]) we expected the time-domain-based algorithm to perform better than the magnitude-based algorithm. One goal of this study was to see if and which subjective methods confirmed this result.

2.1 Magnitude-based algorithm

We utilized a pre-trained LSTM-based embedder network [22] and a separator network containing a combination of ResNet and gated recurrent units (GRUs). The embedder network was pre-trained on the Voxceleb dataset [23] to generate a 256-dimensional speaker embedding. The separator network architecture was inspired by [5], where instead of utilizing convolution neural networks (CNNs) with LSTMs, we utilized a ResNet with GRU layers to estimate a soft mask for target speaker extraction. The separator network consists of 2 ResNet layers, 2 uni-directional GRU layers, and 2 fully connected (FC) layers. Each ResNet layer consists of 2 basic blocks of 2 CNN layers, each followed by batch normalization and ReLU activation. The last ResNet layer is average pooling. The number of nodes utilized for the GRU layers and the first FC layer is fixed to 256, while the last FC layer consists of 257 nodes with a sigmoid activation function. As input, the separator network uses the magnitude of the STFT coefficients of the mixture $y(n)$ computed using a square root Hann window with a frame length of 512 samples and a shift of 256 samples at a sampling frequency of 16 kHz. The separator network was trained using the SI-SNR loss function [3] using the ADAM optimizer with a learning rate of 0.0002 for 150 epochs utilizing the early stopping criteria. The separator network has 3.8 M parameters, while the embedder network has 12.1 M parameters.

2.2 Time-domain-based algorithm

We utilized a recently proposed target speaker extraction system as in [9], which consists of a ResNet-based embedder network and a TCN-Conformer-based separator network. The embedder network generates a 256-dimensional speaker embedding. The separator network consists of a multi-scale speech encoder, 4 stacks of TCN and conformer blocks, and a multi-scale speech decoder. The input and convolutional size of each TCN block are fixed to 512, and the kernel size is fixed to 3. Each conformer block utilizes 8-head attention with a convolutional kernel size of 31. The output size of the feed-forward layer of each conformer block is 4 times the input size. The feed-forward layer is followed by a swish activation and a dropout layer. The algorithm was trained using a weighted combination of multi-scale SI-SNR loss for the separator network and cross-entropy loss for the embedder network using the ADAM optimizer with a learning rate of 0.001 for 150 epochs utilizing the early stopping criteria. The algorithm has a total 12.8 M parameters.

3 Subjective evaluation

3.1 Methods

3.1.1 Participants and stimuli

15 native German-speaking subjects between 19 and 32 years participated in the listening tests. All had normal hearing according to clinical audiometry.
The target speaker stimuli consisted of German matrix sentences uttered by a fixed male speaker from the Oldenburg sentence test [18]. These sentences always contain 5 words in the fixed order *name verb numeral adjective object* (e.g., Peter has eight big chairs). For each word, ten alternatives are available which can be randomly combined to produce syntactically correct, but semantically unpredictable sentences. The speaker embeddings are obtained from 10 s of reference speech of the target speaker chosen from the German Göttingen sentence test [24]. Interfering speech also consisted of matrix sentences uttered by either one or two different speakers. Interferer signals were generated by concatenating several sentences, starting at a random position for each presentation. The interfering speakers were either male (same as the target) or female. The relative level of the target speaker and the interfering speaker(s) was varied to produce different SNRs. To familiarize the participants with the target speaker, participants listened to an example of about 60 s consisting of concatenated target sentences. The sentences were superimposed by interfering speakers as in the experiments, but the SNR was such that target speaker was clearly louder than the interfering speakers to make participants aware of target speaker. During the experiments, stimuli were presented via Sennheiser HD650 headphones in sound-attenuated booths.

3.1.2 Procedures

3 different methods were utilized to assess the performance of both considered algorithms. The methods differ with respect to the outcome measure and the SNR range to which they are applicable.
Paired comparisons were measured for stimuli in which the target speaker was masked by either one or two interfering speakers. The signals were scaled such that the target speaker had the same level (70 dB SPL) as the single or the combined interferers. In each trial, participants listened to two versions of the same stimulus visually marked as intervals A and B. By clicking the intervals they could toggle between the two versions as long as they liked (the stimuli were played back in a loop) and decide in which interval the target speaker was more intelligible [25]. The rating was made on a six-point scale on which either interval could be marked as much easier (A+++ / B+++), clearly easier (A++ / B++), or easier (A+ / B+). There was no middle category (equally easy), i.e., participants had to decide for one interval in each trial. All three versions (unprocessed, magnitude-based, time-domain-based) were compared with each other, and the assignment to intervals A and B was randomized in each trial. For each comparison, 3 repetitions with different target and interferer sentences were made. The data were analyzed with

respect to the percentage of wins.
Speech intelligibility was measured in terms of SRTs for stimuli in which the target speaker was masked by two interfering speakers because for a single interfering speaker SRTs are known to be extremely low (e.g., −14 dB SNR in [26]), i.e., they would have been expected to be in an SNR region where we did not expect the algorithms to work properly nor where typical listening conditions would occur. SRTs were measured using a list of 20 sentences. In each trial, one mix of target speaker and interfering speakers (processed or unprocessed) was presented once. The participants then marked the recognized words on a screen, before proceeding to the next trial. The level of the combined interfering speakers was fixed at 70 dB SPL, while the level of the target speaker was adjusted adaptively depending on the participants' responses in the previous trial. If the participants correctly recognized three or more of the five words, the SNR was decreased, otherwise, the SNR was increased [27]. The adaptive procedure converged to the SRT. Each condition (unprocessed, magnitude-based, time-domain-based) was measured with its own list of 20 sentences and in random order. Before conducting the actual SRT measurements, two training lists of 20 target sentences in stationary noise were measured to reduce training effects [18].
The perceived listening effort was measured using a categorical scaling procedure with fixed SNRs [19]. In each trial, participants heard a mix of target speaker and interfering speaker(s) (processed or unprocessed) and had to rate the perceived effort it took them to understand the target speaker. The rating was made on a 13-point scale ranging from "no effort" (1 effort scaling categorical unit, ESCU, see Fig. 1 in [28]) to "extreme effort" (13 ESCU). A fourteenth category "only interfering speakers" was provided for trials in which participants could not hear the target speaker at all (and hence an assessment of the effort related to listening to this speaker could not be reasonably made). In each trial, the stimulus was played in a loop until participants made their rating, which triggered the next trial. All SNRs and all processing conditions were presented together in random order. Target and interfering speakers were mixed at SNRs between −10 and 15 dB in steps of 5 dB. To avoid large differences in loudness between trials, the overall level of the stimulus was always fixed at 70 dB SPL. Listening effort was measured for stimuli in which the target speaker was masked by either one or two interfering speakers. Each combination of SNR, processing condition, and interferer condition was measured 3 times using different sentences. The median across these repetitions was used as the estimate of each individual's perceived listening effort for that combination.

3.2 Results

3.2.1 Paired comparisons

Figure 1 shows the percentage of wins of paired comparisons between unprocessed stimuli and stimuli processed by each algorithm (top panels), and the direct comparison between both algorithms (bottom panel). The data reveals a very clear preference for the time-domain-based algorithm: In comparison to unprocessed stimuli, 100% of all comparisons favored processed stimuli for a single interferer. For two interferers, the percentage of wins was about 96% (female) and 98% (male). The distribution of ratings indicates that in most cases the processed stimuli

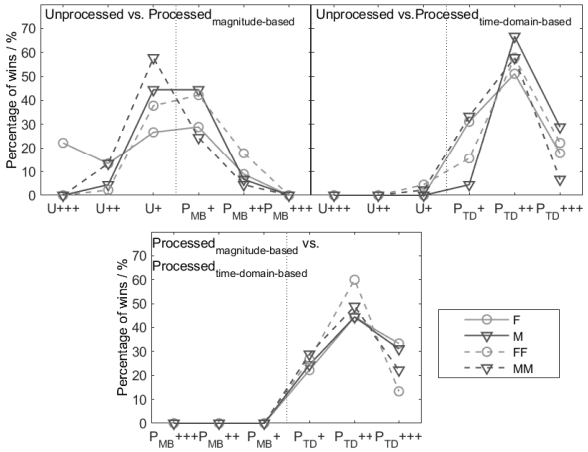

Figure 1: Percentage of wins in the paired comparisons for each pair of the three processing conditions for either a single male or female interferer (F/M) or two interferers (FF/MM).

were perceived as "clearly easier" to understand (P++) than the unprocessed stimuli.
In contrast, the magnitude-based algorithm did not provide a consistent advantage over unprocessed stimuli. Most ratings were given to the middle categories of the rating scale, indicating that listeners were uncertain whether unprocessed or processed stimuli were easier to understand. Consistently, the time-domain-based algorithm was strongly preferred over the magnitude-based algorithm.

3.2.2 Speech intelligibility

Figure 2 shows measured SRTs averaged across participants for 3 processing conditions, where error bars represent standard errors. Statistical significance was tested using a two-way analysis of variance for repeated measures, followed by Bonferroni-corrected t-tests as post-hoc tests. In general, SRTs were negative for all processing conditions (indicating that listeners were able to understand 50% of the target when the energy of the interferers exceeded the energy of the target) and significantly lower for female interferers than for male interferers ($F(1,14)=203.259$, $p<0.001$). The factor processing condition ($F(2,28)=18.833$, $p<0.001$) as well as the interaction between processing condition and interferer sex ($F(2,28)=41.003$, $p<0.001$) also significantly affected the performance. For male interferers, the highest mean SRT achieved by the magnitude-based algorithm was −3.6 dB, while the time-domain-based algorithm improved SRTs (−7.9 dB) compared to unprocessed stimuli (−5.0 dB). For female interferers, the lowest SRT for unprocessed stimuli was −13 dB, while SRTs for processed stimuli were about 3 dB higher for both processed conditions (−9.9 and −10.3 dB for magnitude-based and time-domain-based, respectively).

3.2.3 Perceived listening effort

The top panels of Figure 3 show median listening effort ratings across participants as a function of SNR for one interferer (left) and two interferers (right). Different line styles represent 3 processing conditions (error bars are not shown to increase readability). The bottom panels show the corresponding benefit in listening effort due

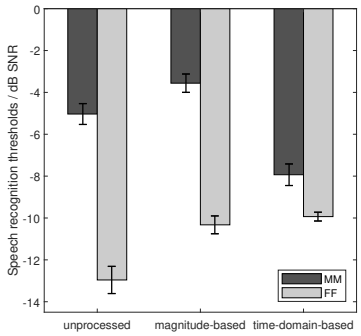

Figure 2: Mean SRTs across participants for the three processing conditions for two interferes (MM, FF). Error bars represent standard errors.

to algorithms relative to the unprocessed condition. For unprocessed stimuli, perceived listening effort decreased systematically with increasing SNR. At the two lowest SNRs, listening effort was higher for two interferers than for one interferer. For magnitude-based, perceived effort was similar to the unprocessed condition for lower SNRs, and considerably higher at high SNRs. In contrast, the time-domain-based algorithm showed a consistent decrease in listening effort compared to the unprocessed condition at all SNRs for both interferer sexes except at −10 dB. The mean benefit due to processing (bottom panels) was the largest at intermediate SNRs and reached more than 4 categories on the 13-point scale for a single interferer. For two interferers, the benefit had a similar pattern but was slightly smaller than for a single interferer. This was confirmed by single-sample t-tests conducted to test if the mean benefits differed significantly from 0. Even at a conservatively corrected significance level of 0.05/24 (there were 24 tests for each interferer condition), the t-tests revealed statistically significant increases in listening effort for the magnitude-based algorithm at SNRs of 10 dB for a single male interferer, and at 10 and 15 dB for a single female interferer. For time-domain-based algorithm, there was a statistically significant reduction in listening effort for SNRs from −5 to 10 dB for both interferer sexes. For two interferers, magnitude-based algorithm significantly increased listening effort at +15 dB (male) and 5 dB (female), while time-domain-based algorithm significantly improved listening effort at 0 and 10 dB (male), and at −5, 0, and 10 dB (female).

4 Discussion

In line with instrumental evaluations, the present data revealed a consistent performance ranking of both considered target speaker extraction algorithms across experimental methods, in that the magnitude-based algorithm did not produce a measurable improvement over the unprocessed condition, while the time-domain-based algorithm showed considerable improvements. The benefit of the time-domain-based algorithm was found both at an SNR of 0 dB (see paired comparisons and listening effort ratings), which is often used in instrumental evaluations of speaker extraction algorithms, as well as at lower and higher SNRs (see listening effort ratings). Strikingly, the SRT measurements revealed that an improvement is even possible with this algorithm for male (i.e., same-sex) interferers, despite the fact that SRTs in the unprocessed

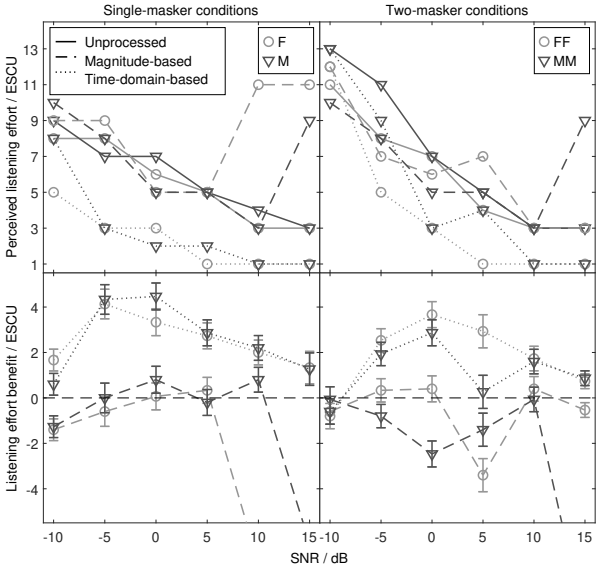

Figure 3: Perceived listening effort (top) and benefit in listening effort relative to unprocessed stimuli (bottom) as a function of SNR for a single interferer (left) and two interferers (right).

condition were as low as −5 dB SNR. Such low SRTs for unprocessed stimuli were also reported in [29, 30] for very similar stimuli and methods. Similarly, these studies also found the effect of sex differences between target and interfering speakers, which is in line with the presented data (SRT decrease by about 8 dB). SRTs in the processed conditions showed that the algorithms did not provide an advantage over unprocessed speech for female interferer, presumably because the SNR was too low (−13 dB) to effectively extract the target voice. The same observation was made for listening effort ratings, which showed a benefit for the time-domain-based algorithm across a large range of SNRs (except for −10 dB for two interferers).
It can be observed that the performance of the considered algorithms does not generalize well when the test SNR conditions deviate from the SNR conditions during training (between 0 and 5 dB). For example, the performance of the time-domain-based algorithm considerably degrades at −10 dB. In addition, especially at high SNRs (10 and 15 dB), the magnitude-based algorithm showed a large increase in listening effort. One possible reason for such poor performance is distortions introduced by the algorithm in the estimated spectra of the target speaker, which makes the extracted signal less intelligible than the unprocessed mixture.

5 Conclusions

We investigated the performance of two different speaker-conditioned target speaker extraction algorithms. All employed subjective measures agreed that a large improvement of target speaker perception can be obtained by the time-domain-based algorithm compared to the unprocessed mixture, even down to low SNRs. In contrast, the magnitude-based algorithm showed no improvement over the unprocessed mixture indicating that the subjective measures replicate the instrumental evaluation results. The measurement methods investigated in this study seem suitable for evaluating speaker extraction algorithms and can be selected according to the SNR range of interest.

References

[1] A. W. Bronkhorst, "The cocktail-party problem revisited: early processing and selection of multi-talker speech," *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, 2015.

[2] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018.

[3] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[4] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Toronto, Canada), pp. 21–25, 2021.

[5] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. Interspeech*, (Graz, Austria), pp. 2728–2732, Sept. 2019.

[6] M. Delcroix, T. Ochiai, K. Žmolíková, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Barcelona, Spain), pp. 691–695, May 2020.

[7] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," in *Proc. Interspeech*, (Shanghai, China), pp. 1406–1410, Oct. 2020.

[8] P. Shen, S. He, and X. Zhang, "Exarn: self-attending RNN for target speaker extraction," *arXiv preprint arXiv:2212.01106*, 2022.

[9] R. Sinha, M. Tammen, C. Rollwage, and S. Doclo, "Speaker-conditioning single-channel target speaker extraction using conformer-based architectures," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, (Bamberg, Germany), pp. 1–5, Sept 2022.

[10] C. Xu, W. Rao, E. S. Chng, and H. Li, "Time-domain speaker extraction network," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, (Sentosa, Singapore), pp. 327–334, IEEE, Dec. 2019.

[11] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. on Graphics*, vol. 37, no. 4, pp. 1–11, 2018.

[12] R. Gu, L. Chen, S.-X. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Neural spatial filter: Target speaker speech separation assisted with directional information.," in *Proc. Interspeech*, (Graz, Austria), pp. 4290–4294, Sept. 2019.

[13] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs P.862," tech. rep., International Telecommunications Union (ITU-T) Recommendation, Feb. 2001.

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

[15] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Brighton, UK), pp. 626–630, May 2019.

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

[17] E. Parizet, "Paired comparison listening tests and circular error rates," *Acta acustica united with Acustica*, vol. 88, no. 4, pp. 594–598, 2002.

[18] K. Wagener, T. Brand, and B. Kollmeier, "Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil III : Evaluation des Oldenburger Satztests (Development and evaluation of a German sentence test Part III : Evaluation of the Oldenburg sentence test)," *Zeitschrift für Audiologie*, vol. 38, no. 3, pp. 86–95, 1999.

[19] J. Rennies, H. Schepker, I. Holube, and B. Kollmeier, "Listening effort and speech intelligibility in listening situations affected by noise and reverberation," *J. Acoust. Soc. Am.*, vol. 136, pp. 2642–2653, 2014.

[20] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Shanghai, China), pp. 31–35, Mar. 2016.

[21] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.

[22] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Calgary, Canada), pp. 4879–4883, Apr. 2018.

[23] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Proc. Interspeech*, (Stockholm, Sweden), pp. 2616–2620, Aug. 2017.

[24] B. Kollmeier and M. Wesselkamp, "Development and evaluation of a german sentence test for objective and subjective speech intelligibility assessment," *J. Acoust. Soc. Am*, vol. 102, no. 4, pp. 2412–2421, 1997.

[25] E. Parizet, N. Hamzaoui, and G. Sabatie, "Comparison of some listening test methods: a case study," *Acta Acustica united with Acustica*, vol. 91, no. 2, pp. 356–364, 2005.

[26] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Release from informational masking by time reversal of native and non-native interfering speech (L)," *J. Acoust. Soc. Am.*, vol. 118, no. 3, pp. 1274–1277, 2005.

[27] T. Brand and V. Hohmann, "An adaptive procedure for categorical loudness scaling," *J. Acoust. Soc. Am.*, vol. 112, no. 4, pp. 1597–1604, 2002.

[28] M. Krueger, M. Schulte, T. Brand, and I. Holube, "Development of an adaptive scaling method for subjective listening effort," *J. Acoust. Soc. Am.*, vol. 141, pp. 4680–4693, jun 2017.

[29] G. Kidd, C. R. Mason, J. Swaminathan, E. Roverud, K. K. Clayton, and V. Best, "Determining the energetic and informational components of speech-on-speech masking," *J. Acoust. Soc. Am.*, vol. 140, pp. 132–144, jul 2016.

[30] J. Rennies, V. Best, E. Roverud, and G. Kidd Jr., "Energetic and informational components of speech-on-speech masking in binaural speech intelligibility and perceived listening effort," *Trends in Hearing*, vol. 23, pp. 1–21, 2019.