INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Intelligibility-enhancing speech modifications – The Hurricane Challenge 2.0

Jan Rennies[1], *Henning Schepker*[2], *Cassia Valentini-Botinhao*[3], *Martin Cooke*[4]

[1]Fraunhofer IDMT and Center of Excellence Hearing4All, Oldenburg, Germany
[2]University of Oldenburg, Signal Processing Group, Oldenburg, Germany
[3]The Centre for Speech Technology Research, University of Edinburgh, UK
[4]Basque Foundation for Science, Bilbao, Spain

jan.rennies@idmt.fraunhofer.de, henning.schepker@uni-oldenburg.de,
cvbotinh@inf.ed.ac.uk, m.cooke@ikerbasque.org

Abstract

Understanding speech played back in noisy and reverberant conditions remains a challenging task. This paper describes the Hurricane Challenge 2.0, the second large-scale evaluation of algorithms aiming to solve the near-end listening enhancement problem. The challenge consisted of modifying German, English, and Spanish speech, which was then evaluated by a total of 187 listeners at three sites. Nine algorithms participated in the challenge. Results indicate a large variability in performance between the algorithms, and that some entries achieved large speech intelligibility benefits. The largest observed benefits corresponded to intensity changes of about 7 dB, which exceeded the results obtained in the previous challenge despite more complex listening conditions. A priori information about the acoustic conditions did not provide a general advantage.

Index Terms: intelligibility, speech enhancement

1. Introduction

In many everyday situations speech is played back to convey information (e.g., public address systems, mobile phones, smart speakers). However, the intended speech signal is often joined by competing sounds in the listening environment or degraded by properties of the transmission channel (e.g. reverberation). One way to maintain high intelligibility under adverse conditions is to increase the intensity of the playback signal to improve the signal-to-noise ratio (SNR). However, this is only possible to a limited degree in practice since output levels that are too high may result in discomfort or overload the playback equipment. Consequently, alternatives are needed to modify the speech signal with the aim of maintaining intelligibility under an equal-level constraint. Various approaches have been proposed to tackle the so-called near-end listening enhancement (NELE) problem. These include modifications of spectral properties (e.g., [1-6]), non-linear amplification such as dynamic range compression (e.g., [7-10]), selective enhancement of certain signal components (e.g., [11-13]) or speech modulations (e.g., [14]), and time-scale modification (e.g., [15-17]. Because these algorithms are typically explored and validated in isolation or compared against different baselines, a large-scale evaluation was initiated in 2012 with a goal of comparing the performance of NELE algorithms using shared data and metrics. This evaluation took place within the Hurricane Challenge [18]. The focus of the Hurricane Challenge was interfering sounds, and two types of maskers were used (stationary speech-shaped noise and a single competing talker). One main result was that large performance differences were observed between algorithms, and that only a few algorithms could provide intelligibility benefits for both types of maskers. Most of the successful algorithms employed dynamic range compression indicating that level-dependent amplification is a powerful approach to this challenging problem. Another important observation was that algorithms that employed a priori knowledge of the maskers did not necessarily perform better than noise-independent algorithms. While the Hurricane Challenge was the first open large-scale NELE algorithm comparison and, as such, provided a number of qualitative and quantitative insights into the effectiveness of speech modification techniques, it was limited in its scope. For example, the interferers employed represented two highly artificial masking environments, while more realistic acoustic conditions such as multitalker environments [19] and reverberation [20] have been shown to be challenging for NELE algorithms [21].

This paper presents the results of the second Hurricane Challenge, in which some of these limitations were addressed. A major extension, in addition to the use of more realistic masking noise, was the inclusion of different degrees of reverberation, allowing for a more general assessment of algorithmic benefit in real rooms. In addition, subjective evaluations were carried out in three different languages, permitting a multilingual comparison of algorithms. Finally, knowledge of the listening conditions was limited in a more realistic way by not providing the exact waveforms of the masking noise, but only a waveform and room impulse responses recorded in the same room. The new challenge therefore tests NELE algorithms in a more realistic way, and allows for more representative assessment of recent techniques in real applications.

2. The Challenge problem

Entrants were provided with a corpus of waveforms including the target speech as well as noise and room impulse responses (RIRs) recorded in the proximity of the listener position. Entrants returned algorithmically-modified target speech signals for each of the three speech corpora. These were then subjected to evaluation by listeners. Entrants had around two months to prepare their modified signals, and made a financial contribution to the cost of listening tests.

2.1. Target speech

The target speech consisted of recordings of matrix sentences uttered by one male speaker without foreign accent in each language (German, American English and European Spanish)

http://dx.doi.org/10.21437/Interspeech.2020-1641

[22]. Matrix sentences have fixed five-word structure (e.g., 'Peter bought eight big chairs'). For each word, ten alternatives are available which allows constructing syntactically correct, but semantically unpredictable sentences. While there are language specific differences (e.g., with respect to the order of object and adjective in Spanish vs. German/English), the general concept as well as the design criteria are highly comparable across tests in different languages [23]. One major advantage of matrix sentences compared to context-rich everyday sentences is that they cannot be easily memorized and, hence, multiple repetitions of the test are possible without being restricted by the number of available test lists. A practical disadvantage is that listeners have to be made familiar with speech material to limit the effect of training.

2.2. Acoustic conditions

For each language, nine conditions were defined including three different reverberation conditions (RIRs) and three different SNRs for each RIR. The acoustic conditions were created from recordings of masking noise and impulse responses. These were used to simulate the speech transmission path in a room with variable room acoustics (Figure 1). In the configuration used for the Challenge, the room had a broadband reverberation time of about 0.8s.

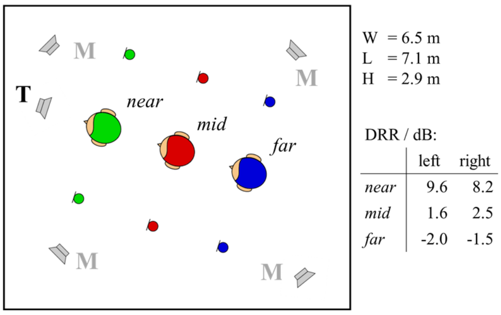

Figure 1: Schematic diagram of the recording setup.

Target speech was presented from the frontal loudspeaker (labelled 'T') and reverberation was varied by changing the distance of the listener (dummy heads in the figure) to the target speaker from 1m ('near') to 2.5m ('mid'), and 4m ('far'), resulting in different direct-to-reverberant ratios. For each position, two sets of impulse responses were recorded: binaural head-related impulse response (BHRIRs) at the listener position, recorded with a head-and-torso simulator, and two single-channel RIRs recorded at positions approximately 1.5m to the left and the right of the listener using omnidirectional microphones (Figure 1). The BHRIRs were used in the listening tests while the single-channel RIRs were provided to Challenge participants. The masking noise (labelled 'M') was cafeteria babble [24] played back from four loudspeakers oriented towards the corners of the room (uncorrelated noise tokens per speaker). Similar to the impulse responses, binaural noise recordings at the listener position were used in the listening tests, while the single-channel recordings at the lateral microphones were provided to the participants. Therefore, the noise signals and RIRs available to the participants were not sample-by-sample equivalents to those used for the subjective listening tests. However, they were recorded in the same acoustic scene.

To create the stimuli used for the evaluation, speech signals were convolved with the binaural RIRs and centrally embedded in segments of the cafeteria noise recordings made for each of the reverberant conditions. Each masker was 2s longer than the corresponding sentence, yielding 1s leading and 1s lagging masker noise, and speech signals were padded with 1s leading and 1s lagging zeros. This allowed a comparison between modifications which produced speech of different lengths, permitting temporal elongation of each sentence by up to 2s. SNRs were set independently for each language and reverberant condition based on pilot experiments, and were selected to correspond to approximately 25% ('low'), 50% ('mid'), and 75% ('high') correctly understood words. SNR was defined as the intensity ratio between the reverberant speech (measured as active speech level [25]) and the masking noise associated with the corresponding sentence (measured as root-mean-square (rms) level). Table 1 summarizes the SNRs used in the Challenge. To create listening test stimuli, speech signals were first convolved with the respective BHRIRs and then rescaled to produce the desired SNR values. To achieve a sufficient degree of reverberation across the whole sentence, prior to convolution three sentences were concatenated of which only the last one was retained for the listening test.

Table 1: *SNRs (in dB) employed in the challenge.*

Language	Rev.	Low SNR	Mid SNR	High SNR
English	Near	-13.0	-8.5	-4.0
	Mid	-11.0	-5.0	1.0
	Far	-10.0	-4.0	2.0
German	Near	-15.0	-12.5	-10.0
	Mid	-13.0	-10.0	-7.0
	Far	-13.0	-9.0	-5.0
Spanish	Near	-17.5	-14.5	-11.5
	Mid	-17.0	-14.0	-11.0
	Far	-18.0	-14.0	-10.0

3. Challenge entries

The following algorithms were submitted to the Challenge. Whether or not they are noise- and reverberation-dependent is indicated in Table 2.

ACO [26]: Sequential combination of modified versions of the algorithms AdaptDRC [19] and Onset-Enhancement [27]. AdaptDRC aims at enhancing high-frequency and low-energy regions of speech when intelligibility is predicted to be low due to additional noise by an estimate of the Speech Intelligibility Index (SII). Onset-Enhancement aims at reducing overlap-masking of speech as well as enhancing its onsets to improve intelligibility in reverberant environments.

ASE [28]: Aims to apply sound engineering knowledge to maximize speech intelligibility while achieving high sound quality without the need for any input parameter besides the target signal itself. ASE divides the signal in six frequency bands. Each band is compressed individually using ad-hoc parameters, with a non-conventional compressor. The signal is then equalized to maximize intelligibility while considering human loudness perception. After the signal is reconstructed, an additional stage of broadband compression is performed. In the (beta) version entered for the Challenge, parameters for compression and equalization were based on expert knowledge, and were fixed for all processed signals.

exactMaxSII: Filters speech signals to maximize the SII. The algorithm is based on an exact solution of the optimization problem, while previous work aimed to maximize approximations of the SII. From a given long-term noise spectrum and SNR, the algorithm computes the optimal speech spectrum that maximizes SII in the given conditions. Mean equalization coefficients are then computed to design a stable fixed equalizer to apply to the speech signal.

DeepSSC-Lomb: Parametric Speaking Style Conversion (SSC) approach based on the training of deep Recurrent Neural Networks on the Lombard-GRID dataset. In addition to standard SSC speech features (fundamental frequency and energy in log scales, Mel-Frequency Cepstral Coefficients) a Continuous Wavelet Transform is used to describe the pitch and energy features at different time scales.

DSSC-L/eMSII: Sequential combination of the two algorithms DeepSSC-Lomb and exactMaxSII.

iMetricGAN [29]: A generator (G) and a discriminator (D) are designed. D tries to accurately predict instrumental intelligibility scores (SIIB [30] and ESTOI [31]) of modified speech, and then guides G to modify input speech in such a way to maximize the predicted intelligibility scores. G receives unmodified speech and noise, and outputs scale factors, which are point-wise multiplied with the unmodified spectrogram to produce the modified one. Modified speech is then re-synthesized by ISTFT. Reverberation is not explicitly accounted for in this approach.

MS500: Speech remains intelligible if its modulation spectrum (MS) resists smearing by the modulation transfer function (MTF) of the environment. If the smeared MS is obtained by multiplying the MTF with the MS of original speech (MS_o), then multiplying MS_o by the inverse MTF yields an optimally resistant MS. However, obtaining the inverse MTF directly is difficult. MS500 modifies MS_o on significant acoustic and modulation frequencies from relations between the smeared and resistant MS and the MTF for intelligibility.

IISPA [32]: The intelligibility-improving signal processing approach (IISPA) was optimized with an automatic-speech-recognition-based model of speech perception using the provided natural speech material. The optimized IISPA parameters were band-pass edge frequencies, spectral slope and curvature, and spectral modulation compression or expansion. Signal analysis was performed based on a log-scaled Mel-spectrogram and applied with an overlap-add method (free source code is available).

SSDRC [9]: Applies noise-independent spectral shaping and dynamic range compression. It was included as an additional baseline because it was shown to provide very good results in the first Hurricane Challenge.

4. Listening tests

Evaluations took place at three sites (Oldenburg, DE; Edinburgh, UK; Vitoria, ES). At each site, listeners native in the respective country's language (German [N=62], English [N=63], or Spanish [N=62]) were recruited and had to pass audiological screening for normal hearing prior to participation. Signals were delivered via headphones (Oldenburg: Sennheiser HD650, Edinburgh: Beyerdynamic DT770, Vitoria: Sennheiser HD380pro). The headphone transfer functions were equalized for a flat response at a dummy head's ear to minimize differences between sites. All measurements took place in a sound-attenuating booth. Each listener heard 2 sentences of each combination of entry (plain speech + 9 entries), SNR and reverberation in a repeated-measures design (180 sentences in total), and entered the recognized words by marking them on a screen which displayed the entire 5x10 word matrix.

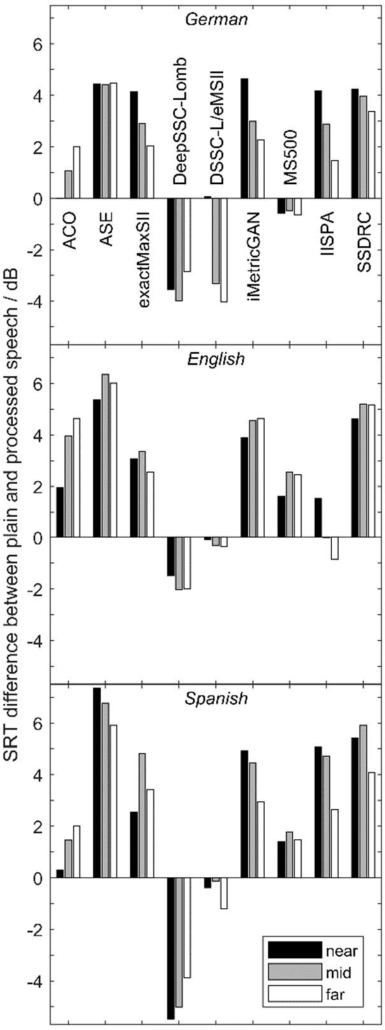

Figure 1: *SRT differences relative to plain speech.*

Prior to the experiments listeners received two lists of 20 sentences as training (not scored). All measurements (including instructions, hearing screening, and training) took place in a single session of approximately 60 min.

Data of one, three, and two listeners for German, English, and Spanish, respectively, were removed based on standard outlier criteria. The mean data of the remaining listeners are summarized in Table 2. To provide an estimate of the effect sizes required for statistical significance, Fisher's least significant difference (LSD) is reported, derived from repeated-measures ANOVAs with factor processing condition, conducted separately for each reverberation and SNR condition. White cells mark a significant change in speech intelligibility in percentage points relative to the plain speech baseline (black font: increase, gray font: decrease). Light gray cells show differences smaller than the LSDs. The best entry for each condition is highlighted in bold face. As additional measure to assess algorithm benefit, the differences in speech recognition threshold (SRT, i.e., the SNR at 50% speech intelligibility) between plain speech and modified speech were derived from psychometric functions fitted to the three data points for each entry, language and reverberation condition (Figure 1).

Table 2: *Differences re. plain baseline scores (given in italics) in percentage points for all sites, entries and conditions.*

		Noise-dep.?	Reverb-dep.?	Reverb near; SNR			Reverb mid; SNR			Reverb far, SNR		
				low	mid	high	low	mid	high	low	mid	high
German	***Plain speech score (LSD)***			*11.1 (6.7)*	*40.8 (6.5)*	*64.9 (6.5)*	*15.4 (6.7)*	*44.6 (6.5)*	*70.3 (6.8)*	*12.3 (6.7)*	*41.0 (6.5)*	*76.7 (6.8)*
	ACO	✓	✓	2.5	-0.8	0.0	9.3	10.7	6.4	9.0	22.6	8.2
	ASE	✗	✗	**50.0**	**45.9**	**29.7**	**43.8**	**44.3**	**26.7**	**31.8**	**44.4**	**20.7**
	exactMaxSII	✓	✗	41.5	30.8	14.4	31.5	13.6	2.6	16.6	14.9	11.5
	DeepSSC-Lomb	✗	✗	-5.7	-31.5	-37.2	-10.2	-31.5	-36.6	-8.9	-23.4	-24.8
	DSSC-L/eMSII	✓	✗	25.6	7.4	-10.5	-2.0	-22.6	-27.2	-5.1	-18.5	-33.6
	iMetricGAN	✓	✗	47.0	33.6	21.8	25.7	29.3	19.2	13.6	23.0	8.7
	MS500	✓	✓	13.1	-2.8	-8.9	2.3	-7.5	-3.4	-4.1	-5.7	-4.8
	IISPA	✓	✓	43.6	31.3	20.3	27.5	21.8	6.6	17.5	12.0	-1.1
	SSDRC	✗	✗	47.0	42.1	29.3	36.4	39.3	21.5	20.8	33.9	16.7
English	***Plain speech score (LSD)***			*7.3 (4.1)*	*18.5 (6.6)*	*50.5 (6.8)*	*13.8 (5.7)*	*43.8 (7.1)*	*73.5 (5.4)*	*18.0 (6.2)*	*42.7 (6.9)*	*75.8 (5.4)*
	ACO	✓	✓	-0.5	8.3	17.8	12.8	26.2	14.0	17.5	27.5	15.8
	ASE	✗	✗	6.8	**42.8**	**40.5**	**27.0**	**42.7**	**23.2**	**22.8**	**42.0**	**18.8**
	exactMaxSII	✓	✗	**10.0**	22.8	18.3	10.3	21.8	12.0	4.0	19.0	9.0
	DeepSSC-Lomb	✗	✗	-4.3	-10.0	-14.2	-4.2	-12.7	-8.7	-7.0	-6.0	-12.3
	DSSC-L/eMSII	✓	✗	2.0	7.8	-1.3	-2.0	-4.3	1.8	-6.3	1.3	-2.7
	iMetricGAN	✓	✗	6.7	27.8	27.5	18.5	26.8	14.8	13.5	34.0	13.5
	MS500	✓	✓	2.3	12.0	11.0	9.3	15.3	8.7	7.5	16.2	6.5
	IISPA	✓	✓	3.7	13.7	9.3	1.0	1.8	-2.5	-2.8	1.5	-9.2
	SSDRC	✗	✗	9.7	31.3	36.7	21.3	31.2	19.5	18.2	34.3	17.7
Spanish	***Plain speech score (LSD)***			*14.8 (6.2)*	*43.2 (6.8)*	*66.3 (5.9)*	*12.7 (6.4)*	*25.5 (6.8)*	*52.5 (5.9)*	*6.8 (6.2)*	*27.5 (6.8)*	*55.0 (5.9)*
	ACO	✓	✓	4.3	-2.7	6.3	1.2	13.0	12.8	0.8	11.0	19.8
	ASE	✗	✗	**58.8**	43.7	**29.0**	**46.7**	**56.5**	**41.2**	**30.2**	**45.7**	**38.2**
	exactMaxSII	✓	✗	23.7	17.2	19.2	32.7	34.2	27.5	22.2	17.7	25.2
	DeepSSC-Lomb	✗	✗	-11.2	-36.8	-46.0	-10.8	-16.3	-34.7	-5.7	-21.7	-32.5
	DSSC-L/eMSII	✓	✗	2.8	-0.2	-6.3	11.2	10.7	-2.2	7.5	-4.5	-6.8
	iMetricGAN	✓	✗	42.2	34.0	23.5	28.0	37.0	23.2	15.2	23.0	15.8
	MS500	✓	✓	17.7	4.5	12.3	7.5	17.8	12.3	4.2	9.0	11.8
	IISPA	✓	✓	41.7	35.5	20.0	30.2	37.5	25.0	17.7	18.5	14.5
	SSDRC	✗	✗	49.0	**44.3**	27.8	39.0	49.2	39.8	15.0	32.0	28.0

5. Discussion

One main result was that entries showed markedly different performance, with four entries providing speech intelligibility benefit in all conditions (ASE, exactMaxSII, iMetricGAN, SSDRC). In some cases, benefits of up to about 60 percentage points (or 7 dB SRT difference) were observed. This is a notable finding because it was shown that cafeteria-like babble can be more challenging for NELE algorithms than less complex stationary maskers even on algorithms that had performed well in the first Hurricane Challenge [18, 21]. Other entries showed significant benefits in some conditions only (ACO, DSSC-L/eMSII, MS500, IISPA) while one had a consistently detrimental effect on intelligibility (DeepSSC-Lomb), which was likely due to clearly audible artefacts. The best entry in all but two of the 27 conditions was ASE, sometimes outperforming the other entries by significant margins and possibly approaching ceiling performance in some cases.

The SSDRC approach, which had been one of the best entries in the previous Challenge, again performed very well, although it had not been developed for speech enhancement in reverberant conditions. Since SSDRC is also noise-independent, this was a striking example of robust speech modification without employing knowledge about the acoustic environment. This is also true in part for iMetricGAN, which made use of the masking noise estimates, but not of the provided RIRs, and still showed good performance in all reverberation conditions. In contrast, ACO (both noise- and reverberation-dependent) performed best in the more reverberant conditions, illustrating that RIR estimates can benefit algorithms aiming to reduce self-masking of speech. The only entries that increased the duration of the target sentences were DSSC-L/eMSII and DeepSSC-Lomb. Their relatively poor performance suggests that time-expansion is not automatically sufficient to obtain good intelligibility if audible artefacts counteract the positive effects expansion may have with respect to robustness in reverberation. The effectiveness of the different entries also appeared to vary with language. While most entries showed smaller benefits for English, ACO showed significantly larger gains. The reasons underlying these language-specific differences are unclear, but it is likely that both language- as well as talker-specific effects played a role [33]. It would be interesting to investigate a larger set of different talkers in each language to learn more about talker-dependency in NELE benefit. Further possible extensions include considering alternative outcome measures. While intelligibility is very important, its measurement often requires employing highly adverse SNRs to avoid ceiling performance already in the baseline condition (like in the present challenge). This may not be representative for many practical applications. For this reason, recent studies have started exploring the effect of NELE processing on listening effort (e.g., [34]), but not yet on a larger scale.

6. Acknowledgements

The authors would like to thank Olympia Simantiraki for helping with the organization of listening tests in Vitoria. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 352015383 – SFB 1330 A 1 and the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska Curie grant agreement no. 675324.

7. References

1. H. Brouckxon, W. Verhelst, and B. Schuymer, “Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments,” *Proceedings of Interspeech*, Brisbane, Australia, pp 557–560, 2008.
2. W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. Petkov, B. Sauert, and P. Vary, “Optimizing speech intelligibility in a noisy environment: A unified view,” *IEEE Signal Process. Mag.* 32, pp. 43–54, 2015.
3. B. Sauert, and P. Vary, “Recursive close-form optimization of spectral audio power allocation for near end listening enhancement,” *Proceedings of the ITG Conference on Speech Communication*, Bochum, Germany, 2010.
4. B. Sauert, and P. Vary, “Near-end listening enhancement in the presence of bandpass noises,” *Proceedings of the ITG Conference on Speech Communication*, Braunschweig, Germany, pp. 195–198, 2012.
5. C. H. Taal, and J. Jensen, “SII-based speech preprocessing for intelligibility improvement in noise,” *Proceedings of Interspeech*, Lyon, France, pp. 3582–3586, 2013.
6. C. H. Taal, R. C. Hendriks, and R. Heusdens, “Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure,” *Comput. Speech Lang.* 28(4), pp. 858–872, 2014.
7. J. C. R. Licklider, and I. Pollack, “Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech,” *J. Acoust. Soc. Am.* 20(1), pp. 42–51, 1948.
8. R. Niederjohn, and J. Grotelueschen, “The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression,” *IEEE Trans. Acoust. Speech* 24(4), pp. 277–282, 1976.
9. T.-C.Zorila, V. Kandia, and Y. Stylianou, “Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression,” *Proceedings of Interspeech*, Portland, OR, pp. 635–638, 2012.
10. T.-C. Zorila, and Y. Stylianou, “On spectral and time domain energy reallocation for speech-in-noise intelligibility enhancement,” *Proceedings of Interspeech*, Singapore, pp. 2050–2054, 2014.
11. T. Arai, N. Hodoshima, and K. Yasu, “Using steady-state suppression to improve speech intelligibility in reverberant environments for elderly listeners,” *IEEE Trans. Audio Speech Lang. Process.* 18(7), pp. 1775–1780, 2010.
12. V. H. M. Ortega, and M. Huckvale, “Automatic cue-enhancement of natural speech for improved intelligibility,” *Speech Hear. Lang.: Work Prog.* 12, pp. 42–56, 2000.
13. M. D. Skowronski, and J. G. Harris, “Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments,” *Speech Commun.* 48(5), 549–558, 2006.
14. A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, and N. Vaughan, “Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments,” *Speech Commun.* 45(2), 101–113, 2005.
15. V. Aubanel, and M. Cooke, “Information-preserving temporal reallocation of speech in the presence of fluctuating maskers,” *Proceedings of Interspeech*, Lyon, France. pp 3592-3596, 2013.
16. Y. Tang, and M. Cooke, “Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints,” *Proceedings of Interspeech*, Florence, Italy, pp. 345–348, 2011.
17. W. Verhelst, “Overlap-add methods for time-scaling of speech,” *Speech Commun.* 30, pp. 207–221, 2000.
18. M. Cooke, C. Mayo, and C. Valentini-Botinhao, “Intelligibility-enhancing speech modifications: The Hurricane challenge,” *Proceedings of Interspeech*, Lyon, France, pp. 3552–3556, 2013.
19. H. Schepker, J. Rennies, and S. Doclo, “Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index,” *J. Acoust. Soc. Am.* 136, pp. 2692–2706, 2015.
20. H. Schepker, D. Hülsmeier, J. Rennies, and S. Doclo, ”Model-based integration of reverberation for noise-adaptive near-end listening enhancement,” *Proceedings of. Interspeech*, Dresden, Germany, pp. 75–79, 2015.
21. C. Chermaz, C. Valentini-Botinhao, H. Schepker, S. King, “Evaluating near end listening enhancement algorithms in realistic environments, *Proceedings of Interspeech*, Graz, Austria, pp. 1373–1377, 2019.
22. S. Hochmuth, B. Kollmeier, and B. Shinn-Cunningham. “The relation between acoustic-phonetic properties and speech intelligibility in noise across languages and talkers,” *Proceedings DAGA 2018*, Munich, Germany, pp. 628–629, 2018.
23. B. Kollmeier, A. Warzybok, S. Hochmuth, M. A. Zokoll, V. Uslar, T. Brand, T., and K. C. Wagener, “The multilingual matrix test: Principles, applications, and comparison across languages: A review,” *Int. J.of Audiol.* 54, pp. 3–16, 2015.
24. H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, “Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses,” *EURASIP Journal on Advances in Signal Processing* 6, 2009.
25. ITU-T P.56. Objective measurement of active speech level, *ITU-T Recommendation*, Geneva, Switzerland, 2012.
26. F. Bederna, H. Schepker, C. Rollwage, S. Doclo, A. Pusch, J. Bitzer , and J. Rennies, “Adaptive compressive onset-enhancement for improved speech intelligibility in noise and reverberation,“ *Proceedings of Interspeech*, Shanghai, China, 2020.
27. J. Grosse and S. van de Par, ”A speech preprocessing method based on overlap-masking reduction to increase intelligibility in reverberant environments,” *J. Audio Eng. Soc.* 65, pp. 31–41, 2017.
28. C. Chermaz and S. King, “A sound engineering approach to near end listening enhancement,“ *Proceedings of Interspeech*, Shanghai, China, 2020.
29. H. Li, S. Fu, Y. Tsao, and J. Yamagishi, “iMetricGAN: Intelligibility Enhancement for Speech-in-Noise using Generative Adversarial Network-based Metric Learning,” *Proceedings of Interspeech*, Shanghai, China, 2020.
30. S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, “An instrumental intelligibility metric based on information theory,” *IEEE Signal Processing Letters* 25, pp. 115–119, 2017.
31. J. Jensen and C. H. Taal, “An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, pp. 2009–2022, 2016.
32. M. Schädler, “Optimization and evaluation of an intelligibility-improving signal processing approach (IISPA) for the Hurricane Challenge 2.0 with FADE,“ *Proceedings of Interspeech*, Shanghai, China, 2020.
33. S. Hochmuth, T. Jürgens, T. Brand, and B. Kollmeier. “Talker- and language-specific effects on speech intelligibility in noise assessed with bilingual talkers: Which language is more robust against noise and reverberation?” *Int. J. Audiol.* 54, pp. 23–34, 2015.
34. J. Rennies, A. Pusch, H. Schepker, and S. Doclo, “Evaluation of a near-end listening enhancement algorithm by combined speech intelligibility and listening effort measurements,” *J. Acoust. Soc. Am.* 144, EL315–EL321, 2018.