INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Evaluating Near End Listening Enhancement Algorithms in Realistic Environments

Carol Chermaz[1], *Cassia Valentini-Botinhao*[1], *Henning Schepker*[2], *Simon King*[1]

[1]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[2]Dept. Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany

c.chermaz@sms.ed.ac.uk

Abstract

Speech playback (e.g., TV, radio, public address) becomes harder to understand in the presence of noise and reverberation. NELE (Near End Listening Enhancement) algorithms can improve intelligibility by modifying the signal before it is played back. Substantial intelligibility improvements have been achieved in the lab for both natural and synthetic speech. However, evidence is still scarce on how these algorithms work under conditions of realistic noise and reverberation.

We present a realistic test platform, featuring two representative everyday scenarios in which speech playback may occur (in the presence of both noise and reverberation): a domestic space (living room) and a public space (cafeteria). The generated stimuli are evaluated by measuring keyword accuracy rates in a listening test with normal hearing subjects.

We use the new platform to compare three state-of-the-art NELE algorithms, employing either noise-adaptive or non-adaptive strategies, and with or without compensation for reverberation.

Index Terms: NELE, Near End Listening Enhancement, realistic noise, reverberation, speech modifications

1. Introduction

Speech playback is very common in everyday life: from television to public announcements in train stations, from laptops to car audio systems. Background noise and reverberation pose an obstacle to the intelligibility of speech. In the case of speech playback, there is a unique opportunity to deploy strategies to reduce this problem: NELE (Near End Listening Enhancement) algorithms can be used to modify the signal *before* it is played by a loudspeaker, in a way that makes it more intelligible for the listener when heard in the presence of noise and reverberation. NELE should not be confused with speech enhancement, which instead attempts to extract the speech signal from a noisy mixture.

Different NELE algorithms have been developed for both natural and synthetic speech in recent years [1, 2, 3, 4, 5, 6], achieving varied degrees of improvement over plain speech. It is common to test these technologies only against additive artificial noise [7, 8]. NELE algorithms are seldom tested against reverberation, except for those which are specifically designed to tackle this problem [9, 10, 11]. It is reasonable to assume that both additive noise and reverberation will be present in most situations in real life, and reverberation alone can hinder communication even in the absence of noise.

In the current study, we start from the conjecture that typical laboratory evaluations of NELE algorithms – which rely on additive artificial noise – might not be a faithful representation of everyday situations, and therefore may yield inaccurate predictions of the performance of NELE algorithms. We propose a realistic test platform with two environments that are representative of everyday scenarios for speech playback: a domestic space (living room) and a public space (cafeteria).

1.1. Previous work

In 2013, a large-scale evaluation study with normal hearing listeners was performed at the University of Edinburgh. The study, known as the Hurricane Challenge [8], compared several NELE algorithms on both natural and synthetic speech. The challenge was an extension of the evaluation described in [7], in which the same noise stimuli and methodologies were used. Intelligibility was scored in terms of WAR (Word Accuracy Rate), i.e., the percentage of correct keywords a subject can recall after listening to a sentence in noise. Intelligibility gains obtained with NELE algorithms were computed in terms of EIC (Equivalent Intensity Change), which is the amount in dB that plain unmodified speech would have to be boosted (or attenuated) in order to achieve the same intelligibility level as the modified speech. A complete description is available in [7].

Two types of additive noise were used in these previous evaluations: SSN (Speech Shaped Noise, which is artificially generated) and CS (Competing Speaker, which is naturally-produced speech from another speaker). These choices represent one steady and one fluctuating noise. The competing speaker of choice was a voice actress speaking sentences in a news-reading style, recorded in a sound studio. SSN was obtained by filtering white noise with the spectral profile of the speech recordings. SSN is meant to represent a situation in which many speakers are present, such as a noisy restaurant. This type of noise is widely used to test algorithms as it is easy to generate and measure, while having some of the spectral properties of more realistic stimuli.

2. Materials and methods

2.1. Design of the realistic environments

Our goal is to create a more realistic test platform, in which the additive noise is more realistic than those above, and in which reverberation is present. We designed two acoustic environments: a small domestic space (the living room) and a large public space (the cafeteria), illustrated in Figure 1. The motivation for this choice is to depict a minimal set of possible scenarios in which speech playback is typically experienced. Since describing the whole spectrum of possibilities would be too broad in scope, these two environments were chosen as representative of opposite characteristics: fluctuating vs steady noise, and short vs long reverberation time.

http://dx.doi.org/10.21437/Interspeech.2019-1800

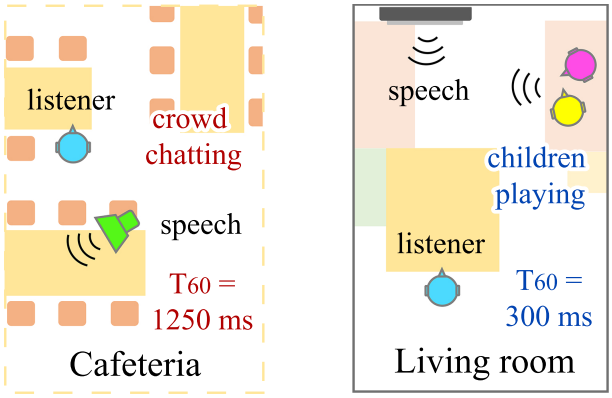

Figure 1: *Schematic representation of the simulated environments; dotted line around the cafeteria represents open boundaries. Drawings are adapted from [12].*

The living room is a small space with a short reverberation time (T60 = 300 ms), and the noise associated with it can be described as fluctuating in nature (it features children playing, house appliances, cars passing by, etc.). For this reason, we assume it is comparable to the CS noise in principle.

The cafeteria, on the other hand, is a wide space with a relatively long reverberation time (T60 = 1250 ms), and it is occupied by a large number of individuals speaking at the same time; this ensemble results in fairly steady-state noise, compared to a single speaker. It is comparable to SSN by definition, yet it contains some variability that SSN fails to capture, such as sources moving around in space and suddenly changing their vocal effort – as well as other types of noise which can be found in a real situation, such as the clinking of cutlery, chairs being moved or doors being shut.

In order to provide a realistic sensation of space and account for sound quality, we selected binaural live noise recordings from The University of Oldenburg's HRIR (Head-Related Impulse Response) database [12] for the cafeteria and from The University of Sheffield's CHiME corpus [13] for the living room. Stimuli were analyzed and recording artifacts were removed, while leaving intact the original spectral characteristics and dynamic range of the HATS (Head and Torso Simulator) recordings. Binaural impulse responses were also taken from the HRIR database ("Office II" for the living room and "Cafeteria" for that scenario) in order to create the reverberant speech stimuli. The position of the speech source in respect to the listener can be seen in Figure 1.

All signals were sampled at 48 kHz with a bit depth of 24 bits.

2.2. Selecting the NELE algorithms

Two state-of-the-art NELE algorithms, SSDRC [4] and AdaptDRC [5, 6], were chosen based on the results they achieved in the Hurricane Challenge [8], where both were among the most effective algorithms for natural speech. Both SSDRC and AdaptDRC operate under an equal-power constraint, i.e., signal power before and after processing must be the same. While SSDRC is only speech-dependent, AdaptDRC performs speech- and noise-dependent processing. Specifically, SSDRC (which was presented in the uwSSDRC variant in [8]) performs speech energy reallocation over both frequency and time. The algorithm performs formant enhancement, boosts the energy in the 1-4 KHz range (spectral shaping) and subsequently the speech energy is reallocated in time via fixed broadband dynamic range compression. Perceptually, SSDRC substantially reduces speech naturalness, although this variable was not measured in the Hurricane Challenge – nor in the current study.

AdaptDRC performs time- and frequency-dependent amplification and dynamic range compression. In contrast to SSDRC, the algorithm is designed to preserve speech naturalness as far possible, whilst at the same time applying modifications whenever intelligibility is not guaranteed, aiming at a fine balance between intelligibility and sound quality. The time- and frequency-dependent amplification is controlled by an estimate of the SII (Speech Intelligibility Index) and applies a uniform distribution of speech power across frequency bands when predicted SII is low, but no processing when predicted SII is high. The time- and frequency-dependent dynamic range compression stage is controlled by the SNR (Signal-to-Noise Ratio), applying maximum compression in situations of low SNR and no compression at high SNRs.

In addition to these two algorithms, the OE (Overlap Masking Reduction and Onset Enhancement) algorithm [11] was chosen as a representative method aimed solely at tackling reverberation. OE is designed to increase the consonant-vowel power ratio to reduce the amount of self-masking of speech. To this end the impulse response of the reverberant environment is assumed to be known, and a frequency-dependent direct-to-reverberant ratio of continuous speech (DRRs) is computed and limited to a maximum value of 25 dB. This DRRs controls the amount of amplification: periods with high DRR are assumed to be consonants and are therefore enhanced, whilst parts with low DRRs are assumed to be vowels and are reduced. In order to fulfill the equal-power constraint, the speech is rescaled after processing. The algorithm was applied to the output of AdaptDRC; the combination of AdaptDRC + OE will be denoted by ADOE in the rest of the paper.

2.3. Listening test design

In order to make the results of this study comparable with those from the Hurricane Challenge, the same corpus for the target speaker and same presentation method were used. Speech stimuli were taken from a recording of the Harvard sentences [14] (as used in [7] and [8]; data are available at https://doi.org/10.7488/ds/2482), which are meant to be phonemically balanced, and are characterized by a relatively low semantic predictability. Sentences were trimmed to have 0.5 s of silence before and after, and were convolved with the impulse responses to simulate reverberation.

Noise stimuli were kept at a fixed level, while the reverberant speech was scaled to achieved the desired SNR. SNR was calculated as 10 times the logarithm of the ratio between the sum of the squared samples of reverberant speech and the sum of the squared samples of noise, taken over the interval where speech is active. The noise snippets were extracted from the noise recordings at random in order to match the length of each sentence. Headphone output was calibrated to 75 dBA for the cafeteria noise and 65 dBA for the living room noise (based on average ambient noise data found in literature). Signal mixtures were presented via Beyerdynamic DT 770 headphones in sound treated booths; listeners had to type onto a keyboard what they had heard.

All participants in the listening tests were normal-hearing native speakers of British English (mean age = 23 years). Subjects were mainly recruited via the University career website

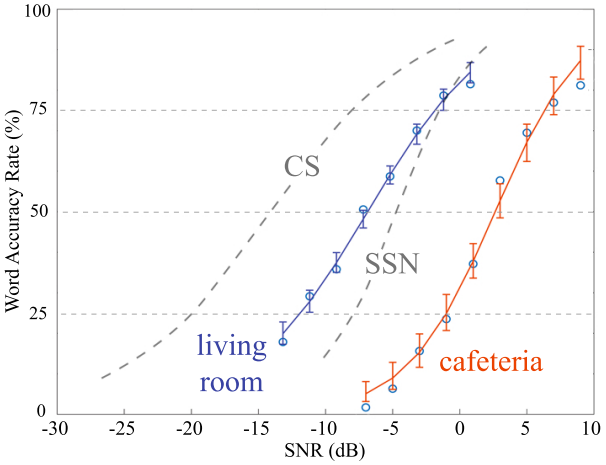

Figure 2: *Comparison between the psychometric curves for plain speech in [7] and present study. 95 % confidence intervals are shown for Matlab's glmfit on SNR data points.*

and all were paid for participation. Prior to the listening test, subjects were screened for hearing loss by means of a Pure Tone Audiometry. Criteria for exclusion were compliant with the definition of hearing impairment of the World Health Organization, i.e. 26 dB or greater hearing threshold (in dB HL) averaged at frequencies 0.5, 1, 2, 4 kHz - in one or both ears.

3. Results

3.1. Psychometric curves

Two calibration studies with respectively N=24 and N=30 listeners were run in order to find the psychometric curves for plain (i.e., unmodified by a NELE algorithm) speech in the two realistic scenarios (Figure 2). For a detailed description of the procedure, the reader can refer to [7], as the methodology is the same (but with different stimuli and SNRs).

A comparison with the data reported in [7] for CS and SSN noise can be seen in Figure 2. The psychometric curves found in the calibration tests suggest the need for higher SNRs, in respect to artificial noise, in order to yield the same intelligibility levels for plain speech in the realistic scenarios. This is true for both the fluctuating noise and the stationary noise condition.

3.2. Comparing NELE algorithms

N=34 listeners participated in the main listening experiment. Sentences were processed with the three algorithms in both realistic noise types, yielding therefore 24 different conditions: 2 types of noise x 4 speech types (3 modified + 1 plain) x 3 SNRs. Stimuli were scaled and added to the appropriate binaurally-recorded noise at three different SNRs, in order to yield close to 25, 50 and 75 % WAR (for plain speech). The SNRs were determined using the psychometric curves from above.

Results from the main experiment are reported in Figure 3. Speech intelligibility is reported in terms of WAR and the corresponding EIC. All algorithms provided intelligibility gains across all conditions. In the cafeteria, larger gains were found in lower SNR conditions, in line with the data reported in [7] and [8]. An opposite trend can instead be observed in the living room for AdaptDRC and ADOE, while SSDRC performed the best at the medium SNR. SSDRC globally obtained the highest scores, with AdaptDRC and ADOE achieving a higher EIC only in the living room at the high SNR. The largest inter-algorithm differences can be observed in the cafeteria/stationary scenario.

As opposed to the findings in [7] and [8], all the algorithms provided more benefit in the fluctuating noise condition, with SSDRC achieving a 5.1 dB EIC gain at Mid SNR, boosting intelligibility from 53.8% to 75.4%. It must be noted that the relationship between WAR and EIC is non linear - it is defined by the psychometric curves in Figure 2.

The addition of OE to AdaptDRC did not appear to change significantly the performance of the latter. ADOE scores were slightly higher in the living room than in the cafeteria, where AdaptDRC performed better in stand-alone mode.

4. Discussion

The SNR differences in the psychometric curves for plain speech between artificial and realistic noise conditions might be explained by the more complex nature of the latter. Both the living room and the cafeteria present stationary *and* fluctuating noise, besides reverberation. In particular, the living room – where the biggest difference can be seen at a low SNR – there are elements of stationary noise (e.g. home appliances), which fill in the glimpses between voice events from the competing speakers. A difference in the spectral profile of the noise stimuli should also be accounted for. A qualitative analysis shows a concentration of energy in the lower frequencies ($<$ 2 kHz) for the CS used in [8] and [7], which might be explained by the proximity effect deriving from a studio recording (with the speaker being close to the microphone in a sound treated room), as opposed to a more natural setting where speakers and noise sources are more distant and affected by the IR of the environment. The IR of the living room, in fact, acts like a high-pass filter. In both the Hurricane Challenge and the present study, SNRs were computed on the raw signals, whereas the differences found in the psychometric curves suggest that it might be worth considering A-weighting (or another perceptually-motivated weighting) before that computation.

In the main listening test, all algorithms provided intelligibility gains, showing that both SSDRC and AdaptDRC have potential in real applications where reverberation is present. It should be noted that our results in terms of EIC are not directly comparable to [8] nor [7], as the psychometric curves (on which the calculation of EIC is based) are inherently different; only trends in the data can be compared, with due considerations on the differences among the studies (with the presence of reverberation being an important factor). The performance of SSDRC and AdaptDRC do not follow the same trends across conditions, yielding unexpected results in the living room, where the highest gains were achieved at the highest SNRs. The reason for this might be found in the complexity of fluctuating noise and reverberation, but informational masking should also be accounted for (especially at a low SNR) as the living room noise features intelligible voices of children.

The ADOE combination gave mixed results, providing most benefit in the living room. This might be explained by the different DRR in the two simulated environments; the source of the speech is closer to the listener in the cafeteria, even though the reverberation time is longer and the speech source is positioned at an unfavourable azimuth. SSDRC provided the largest EIC gains, suggesting that a noise-independent approach can be successful across different scenarios. However, the naturalness of speech is compromised by the modification, which may render it unsuited for long listening periods [15]. The Adapt-

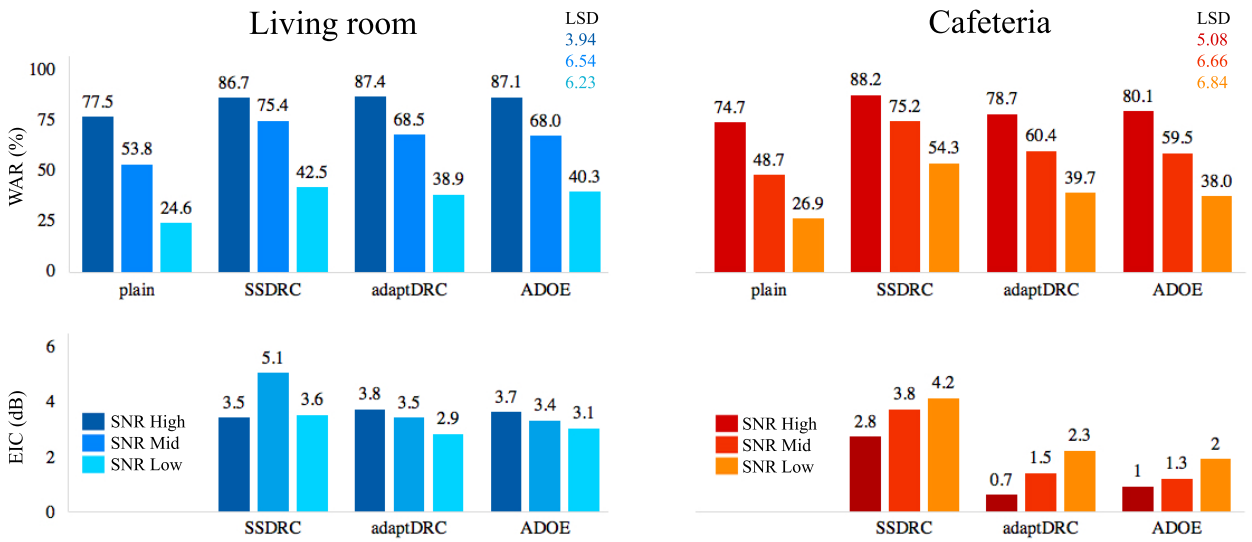

Figure 3: *WAR and EIC for the different algorithms, at each SNR, in the two different noise scenarios. ADOE = AdaptDRC + OE; LSD = Fisher's Least Significant Difference.*

DRC algorithm aims at a compromise between intelligibility and speech naturalness; notwithstanding smaller EIC gains, this strategy might be better suited for long listening periods. A valuable extension to the present study would be a subjective evaluation of perceived quality / naturalness / listening comfort, as well as a subjective and objective measurement of the cognitive load caused by listening to modified speech in realistic noise [16].

Another interesting aspect is the computational load and the intrusiveness of the algorithms. While SSDRC depends only on the speech signal – and therefore can operate in a "blind" manner against any type of noise – AdaptDRC and OE require the noise and the impulse response (or estimates of these) respectively. It is clear that SSDRC can be used in real time, as the stimuli can be processed beforehand; in order to use AdaptDRC or OE in real time, additional resources are needed – not just computational, but also additional hardware to capture the noise or measure the impulse response, such as a microphone in the environment.

5. Conclusions

In this study we tested three state-of-the-art NELE algorithms applied to natural speech, which was then presented in realistic noise and reverberation. We used binaural live noise recordings and impulse responses to create two representative acoustic scenarios: a large crowded space (the cafeteria: stationary noise and long reverberation time) and a small domestic place (the living room: fluctuating noise and short reverberation time).

We ran two calibration listening tests with normal-hearing listeners in order to find the psychometric curves for plain speech in these noise scenarios. This revealed the need for higher SNRs in comparison to synthetic noise [7] in order to achieve the same speech intelligibility levels. We used the data obtained from these tests in order to define the SNRs for the main test, which featured a noise-independent algorithm (SSDRC), a noise-dependent one (AdaptDRC) and ADOE, a combination of AdaptDRC with OE in order to tackle reverberation.

All algorithms improved intelligibility in all conditions; higher gains in EIC were found at lower SNRs in the cafeteria and at higher SNRs in the living room. The noise-independent SSDRC method provided the larger gains overall; the noise-dependent AdaptDRC strategy has a more conservative approach, providing less EIC gain but trying to preserve the naturalness of speech. ADOE performed unexpectedly better in the domestic scenario with a short reverberation time, suggesting the algorithm is quite sensitive to different impulse responses.

Given the higher scores of the non-adaptive, low-computational cost strategy, it is tempting to conclude this as the best solution in all scenarios; however, as already mentioned above, an evaluation of listener *preference* is crucial before this conclusion can be drawn. A useful extension to our study would be to evaluate perceived sound quality and listening comfort, as well as to measure the cognitive load associated with listening to modified speech in noise.

Realistic noise scenarios are difficult to reproduce and control; however, we suggest that their use can provide a critical insight into the expected performance of NELE algorithms – and possibly other technologies – in real applications, as the complexity of real acoustic scenes cannot be captured by lab-controlled noise.

6. Acknowledgements

The authors would like to thank Nina Diviza for her assistance in running the listening experiments; Jan Rennies, Volker Hohmann, Hendrik Kayser and Jon Barker for valuable advice and providing the stimuli; Jason Taylor for help in defining the SNRs for the pilot study; Julie-Anne Meanie for help in recruiting participants.

This project has received funding from the EU's H2020 research and innovation programme under the MSCA GA 67532 (the ENRICH network: www.enrich-etn.eu).

This work was supported by the Deutsche Forschungsgesellschaft (DFG, German Research Foundation) - Project ID 352015383 SFB 1330 C1 and Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project ID 390895286 - EXC 2177/1.

Multimedia files from the experiment are available at http://homepages.inf.ed.ac.uk/s1758351/NELE_RE.html

7. References

[1] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proc. ICASSP*, vol. 1, May 2006, pp. I–I.

[2] C. H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 225–228, March 2013.

[3] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Mel cepstral coefficient modification based on the glimpse proportion measure for improving the intelligibility of hmm-generated synthetic speech in noise," in *Proc. Interspeech*, 2012, pp. 631–634.

[4] T. C. Zorilă, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech*, Portland, USA, September 2012.

[5] H. Schepker, J. Rennies, and S. Doclo, "Improving speech intelligibility in noise by sii-dependent preprocessing using frequency-dependent amplification and dynamic range compression." in *Proc. Interspeech*, Lyon, France, 2013, pp. 3577–3581.

[6] ——, "Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 138, no. 5, pp. 2692–2706, November 2015.

[7] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[8] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proc. Interspeech*, Lyon, France, August 2013.

[9] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto, and T. Kitamura, "Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments," *Acoustical science and technology*, vol. 23, no. 4, pp. 229–232, 2002.

[10] P. N. Petkov and Y. Stylianou, "Adaptive gain control for enhanced speech intelligibility under reverberation," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1434–1438, Oct 2016.

[11] 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*, vol. 65, no. 1/2, pp. 31–41, 2017.

[12] 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*, vol. 2009, p. 6, 2009.

[13] H. Christensen, J. Barker, N. Ma, and P. D. Green, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," in *Proc. Interspeech*, Chiba, Japan, 2010.

[14] IEEE, "IEEE recommended practice for speech quality measurement," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225 – 246, 1969.

[15] Y. Tang, C. Arnold, and T. Cox, "A study on the relationship between the intelligibility and quality of algorithmically-modified speech for normal hearing listeners," *Journal of Otorhinolaryngology, Hearing and Balance Medicine*, vol. 1, no. 1, p. 5, 2018.

[16] J. Rennies, A. Pusch, H. Schepker, and S. Doclo, "Evaluation of a near-end listening enhancement algorithm by combined speech intellgibility and listening effort measurements," *J. Acoust. Soc. Am.*, vol. 144, no. 4, pp. EL315–EL321, October 2018.