PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

[Near end listening enhancement 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, c.chermaz@sms.ed.ac.uk
[(2)]Dept. Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany

Abstract

Speech playback is harder to understand in noise. Near End Listening Enhancement algorithms try to overcome the problem by enhancing the speech signal before it is played by a device. Different strategies have been tried, achieving variable degrees of success in specific noise conditions. Such technologies, however, are often tested in artificial settings - with controlled noise sources and no reverberation. The purpose of this study is to compare a set of state-of-the-art algorithms based on different approaches (adaptive vs non-adaptive, with or without a compensation for reverberation) in simulated real- life scenarios. Binaural noise recordings and impulse responses of real environments have been used to create two representative scenarios in which speech playback may occur, namely a domestic and a public space. A preliminary study with N=24 subjects revealed the need for higher SNRs in the realistic settings (in comparison to controlled noise conditions) in order to obtain the same levels of intelligibility for plain speech. The goal of the main study is to assess the impact of noise adaptivity and reverberation awareness in realistic scenarios, in order to better understand how to make speech playback more robust to noise in real-life situations.

Keywords: Near-end listening enhancement, Realistic Scenarios, Speech intelligibility

1 INTRODUCTION

1.1 Improving speech playback

Speech playback refers to signals played by a device over loudspeakers or headphones, e.g. TV, radio or PA systems. When played over the air, speech sounds becomes mixed with background noise; moreover, reverberation is present in enclosed spaces, posing a hindrance to communication even in the absence of noise. In recent years a range of NELE (Near End Listening Enhancement) algorithms have been proposed in order to make speech playback clearer for the listener. Many of these methods are inspired by behaviours observed in nature, i.e. the ways humans change their speaking style when exposed to noise - such as shifting the vocal effort towards higher frequencies and increasing the consonant/vowel ratio (a phenomenon known as the "Lombard effect").

1.2 State of the art NELE algorithms

In this study we chose to test SSDRC (Spectral Shaping and Dynamic Range Compression) [8] and AdaptDRC [6], which proved to be among the best strategies for natural speech in a large scale NELE evaluation known as "The Hurricane Challenge" [2]. Both algorithms operate on an equal-power-before-and-after processing constraint, but manage to improve intelligibility by relocating the energy in the signal to the regions – in frequency and time – that are most relevant for speech comprehension. While AdaptDRC is aware of the noise and modifies the signal accordingly (applying no modification when there is no noise), SSDRC adapts to the voice of the speaker (as does AdaptDRC) but does not differentiate among different types of noise. Neither of the algorithms accounts for reverberation, nor does any of those analysed in [2], as the problem is typically treated separately in NELE technologies. A common strategy consists in steady state suppression, as vowels (the steady states) ”linger on” in reverberation, and therefore ”smear” signal. The OE (Overlap Masking Reduction and Onset Enhancement)[4] algorithm follows this principle and operates a dynamic amplification control based on the direct-to-reverberant ratio of the speech signal. The impulse response of the environment is assumed to

be known. In this study, the OE algorithm was added to a processing chain after AdaptDRC, creating hence a third algorithm that will be denoted by ADOE.

1.3 Motivation of the study: a realistic test platform

NELE technologies are often tested against artificial noise; furthermore, they are typically not tested against reverberation - besides those algorithms which are specifically designed to tackle this problem. In [2] algorithms were tested against SSN (Speech Shaped Noise) and a competing speaker, whose voice was recorded in a studio by a professional actor. Such signals represent respectively a source of stationary and of fluctuating noise. In this study we take a step towards real-world scenarios, by testing NELE algorithms in simulated realistic environments. Artificial noise may be easy to produce and control, but we argue it fails to represent the complex nature of real acoustic environments – potentially leading to inaccurate predictions on the performance of the technologies at test. For this reason, we have simulated a representative set of scenarios in which speech playback may occur: one domestic space ("the living room") and one public space ("the cafeteria"), which feature recordings of real noise and impulse responses of real spaces.

2 MATERIALS AND METHODS

2.1 Speech, noise and reverberation

Binaural noise recordings from the The University of Oldenburg's HRIR (Head-Related Impulse Response) database [5] were used in order to simulate the cafeteria, while stimuli from The University of Sheffield's CHiME corpus [1] were used to create the living room. Impulse responses were taken also from [5]. The cafeteria is a large public space with a relatively long impulse response; noise recordings were were originally performed at a time in which the place was very crowded, obtaining a relatively stationary signal - which is comparable by definition to SSN. The living room is a small domestic place, with different sources of noise which are active at different times: children playing, appliances from the kitchen, footsteps, etc. The noise can be defined as fluctuating and compared, in principle, to the competing speaker used in [2]. Speech material was taken from a recording of the Harvard sentences (https://doi.org/10.7488/ds/2482), which are phonemically balanced and are characterised by a relatively low semantic predictability. Sentences where processed with the aforementioned NELE algorithms and convolved with impulse responses from [5] in order to simulate speech sources located in a different position in the room in respect to listener. A schematic representation can be seen in Figure 1.

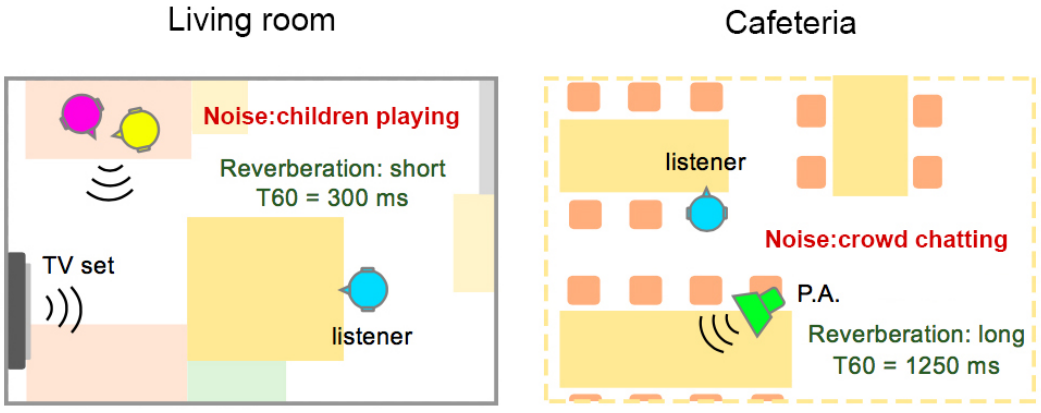

Figure 1. Schematic structure of the simulated environments; dotted line around the cafeteria indicates the space is wider than represented. Images are adapted from [5].

2.2 Calculating the SNRs

SNR was calculated as 10 log (speech power/noise power), where power is represented by the sum of squared samples. Noise snippets were selected at random from the long recordings; endpoints for the sum were defined by the start and end of each sentence.

In a real-world situation, the relative power of speech playback at the position of the listener can be measured with an SPL meter; the measurement, however, will be performed on the reverberant signal, as it will have travelled through the air and will be affected by room acoustics. For this reason, the power of the speech signal was measured on the convolved signal. As the impulse response will change the frequency profile of the signal, special care needs to be taken when performing multiple stages of processing. In the main test, speech stimuli were first processed with the NELE algorithms (as this operation must be performed before the signal is played by a loudspeaker), then convolved with the impulse responses and eventually scaled in order to achieve the desired SNR. We calculated the amount of gain needed for reverberant plain speech in order to reach the desired power over noise; we then applied the same amount of gain to the modified signals – which had the same RMS as plain speech by definition. Noise stimuli were kept at a fixed level. Playback levels were calibrated at 75 dBA (over headphones) for the cafeteria and 65 dBA for the living room. The speech + noise mixtures were presented to normal hearing listeners over Beyerdynamic 770 headphones in sound treated booths.

3 RESULTS AND CONCLUSIONS

3.1 Preliminary tests

Two preliminary listening tests (with respectively N=24 and N=30 normal hearing listeners) were run in order to find the psychometric curves for plain speech in the two realistic scenarios. From this data we calculated the SNRs required for 25, 50 and 75% intelligibility in terms of WAR, which were used in the main test as high, medium and low SNR. We compared the curves for the cafeteria and the living room to those found for SSN and competing speaker in [3]. Our results suggest that higher SNRs are needed in realistic scenarios in comparison to lab controlled noise – up to more than 8 dBs difference for 75% WAR in the SSN vs cafeteria case.

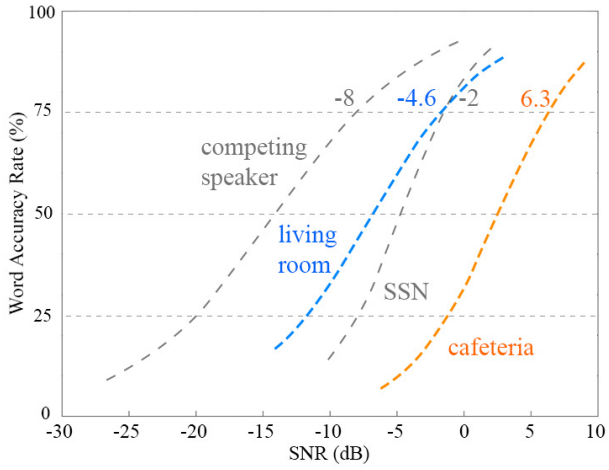

Figure 2. Comparison between the psychometric curves for unmodified speech in [3] and present study. Competing speaker is compared to living room as fluctuating noise, while SSN is compared to cafeteria as stationary noise.

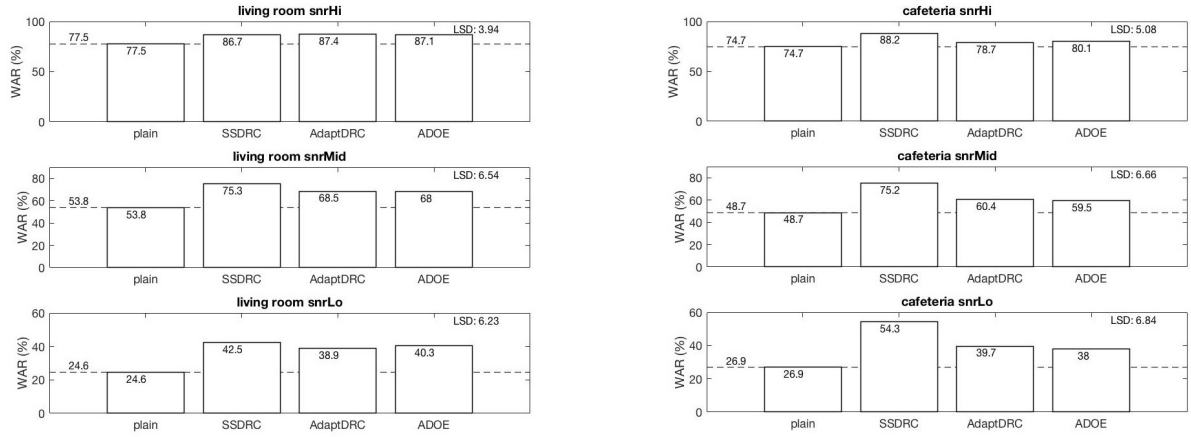

Figure 3. WAR scores for the different algorithms, at each SNR, in the two simulated environments. LSD = Fisher's Least Significant Difference.

3.2 Main test

Results from the main test suggest that different approaches might be best suited for different scenarios. While in the cafeteria scene the higher scores of SSDRC are clear, in the living room scene algorithms offer a comparable performance. No significant improvement was obtained with the addition of OE to AdaptDRC. In almost all conditions SSDRC outperformed the other modifications, suggesting that a blind approach can be very efficient while computationally convenient and non-intrusive, as SSDRC does not require any noise or impulse response measurement. On the other hand, in this study modifications have not been rated for sound quality, which is an important factor for listening comfort in case of long playback times [7]. In SSDRC intelligibility is boosted at the cost of unnatural sounding speech, while AdaptDRC tries to conserve the naturalness of sound at the cost of lower scores in terms of WAR. An important addition to this study would be a subjective and an objective measure of listening comfort in addition to intelligibility measurements.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] 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.

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

[3] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang. Evaluating the intelli-

gibility benefit of speech modifications in known noise conditions. *Speech Communication*, 55(4):572–585, 2013.

[4] 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(1/2):31–41, 2017.

[5] 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*, 2009:6, 2009.

[6] 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*, pages 3577–3581, Lyon, France, 2013.

[7] 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*, 1(1):5, 2018.

[8] 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.