PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Benefits of the WaveNet-Based Speech Intelligibility Enhancement for Normal and Hearing Impaired Listeners

Muhammed Shifas PV[1], Carol Chermaz[2], Theognosia Chimona[3],Vassilis Tsiaras[1],Yannis Stylianou[1]

[1]Speech Signal Processing Lab (SSPL), University of Crete, Greece, shifaspv@csd.uoc.gr
[2]The Centre for Speech Technology Research, The University of Edinburgh, UK
[3]ENT consultant, General Hospital of Chania, Greece

Abstract

Speech perception becomes challenging in adverse listening conditions, hence increasing mental effort. Modification of recorded clean speech with the goal of increasing its intelligibility is one approach for improving listening experience. Recently, we suggested a data-driven WaveNet-like speech intelligibility enhancement method which is based on the Spectral Shaping and Dynamic Range Compressions (SSDRC) approach. Both approaches achieve intelligibility gains by relocating energy from high sonorant to less sonorant portions of speech. In this study we have assessed the performance of the WaveNet intelligibility enhancer for both normal and hearing impaired participants using formal listening tests, in terms of intelligibility and sound quality. Intelligibility was measured as a percentage of correct words recalled (CWR) and quality was assessed via a comparative mean opinion score (CMOS). Furthermore, we compared performance among native and non-native listeners using English and Greek stimuli. We observed that modified speech was significantly more intelligible than plain across all listener groups. Hearing impaired candidates rated modified speech 20% higher on average (on a CMOS scale) than normal hearing subjects. The positive impact for the hearing impaired listeners might be attributed to the reallocation of energy to perceptually relevant frequency bands in which the hearing impaired are less sensitive.

Keywords: Speech intelligibility, listening enhancement, WaveNet approach

1 INTRODUCTION

Modifying plain speech signals with the aim of improving their intelligibility in noise is the goal of near-end listening enhancement (NELE) applications. Modifications are often inspired by natural changes in the human articulatory mechanism when speaking in noise, known as the Lombard effect. Studies have shown the intelligibility gains of Lombard speech over speech produced in quiet; mainly, Lombard speech presents more energy in the higher frequency bands, and is hence characterized by a reduced spectral tilt. Some speech modification models try to capture these structural properties of Lombard speech in order to improve intelligibility.

Based on this idea, different automated signal processing approaches have been developed, e.g. Spectral Shaping and Dynamic Range Compression (SSDRC) [1] and Adaptive Dynamic range compression (AdaptDRC) [2]. AdaptDRC presents a noise-adaptive approach, which adjusts the dynamic range compression of speech depending on the masker signal. In contrast, SSDRC is a masker-independent method: the same processing is performed irrespective of the competing noise. Both these models have proven their ability to boost speech intelligibility in noise. Recent studies have shown the benefits of such speech modifications also in terms of listening effort [3].

Given the proven benefits of speech modifications, we have recently proposed a neural approach for this task [4]: a WaveNet neural model trained to obtain an intelligibility gain equivalent to the SSDRC method. For this reason, our model was named WaveNet-Based SSDRC (wSSDRC). Besides being equally efficient as SSDRC - as shown by a subjective intelligibility evaluation - the neural approach holds the potential of being extended to intelligibility enhancement of real-life speech. Such technology would detect the speech components in a noisy mixture and enhance them with signal modifications that are used in NELE methods, but applied to a different problem - a noisy mixture.

In this work, we assess the intelligibility and quality benefits of wSSDRC through formal listening tests. Evaluations of the model are made on two different language groups: English and Greek. The Greek group includes both normal hearing (NH) and hearing impaired (HI) listeners, while the English group includes NH subjects only. Additionally, in order to examine the different effects of speech modifications for native and non-native speakers of British English, listening tests were carried out on native and non-native groups using the English dataset. The performance of the proposed wSSDRC model is compared throughout conditions with both the SSDRC method and plain unprocessed speech.

2 Speech and masker materials

The Hurricane Challenge natural speech corpus (https://doi.org/10.7488/ds/2482) [5]) was used as speech material for the English listening tests. The corpus consists of a recording of the Harvard Sentences spoken by a British male actor. The Greek sentences were downloaded from the ILSAP database [1]. The dataset is comprised mainly of weather forecasts recorded in quiet. The speech types used in this study are detailed below:

Plain: Unprocessed plain speech selected from the above mentioned databases.
SSDRC-modified speech: The SSDRC modification algorithm [1] is being applied to plain speech to enhance its intelligibility.
wSSDRC-modified speech: This is the proposed neural-based wSSDRC model to be evaluated. As it is a data-driven approach, samples are needed in order to train the model for a given task. Since the model goal was to achieve the intelligibility gains of SSDRC, it was trained on the SSDRC-modified speech, which was set as a target for each plain input. We have trained two separate models for the Greek and English datasets respectively, since the model is language dependent.

Two types of noise are considered for this study; Speech Shaped Noise (SSN) and a Competing speaker (CS), as used in the Hurricane Challenge [5]. The mixing of noise and speech is done at three different SNR levels, corresponding to Low (25%), Medium (50%) and High (75%) intelligibility, in terms of correct words recalled (CWR). As SNR vary across listener groups, exact values are provided in the results section for each group. The quality of processed speech in comparison to plain was assessed in terms of CMOS.

3 RESULTS

3.1 Different Listening Groups

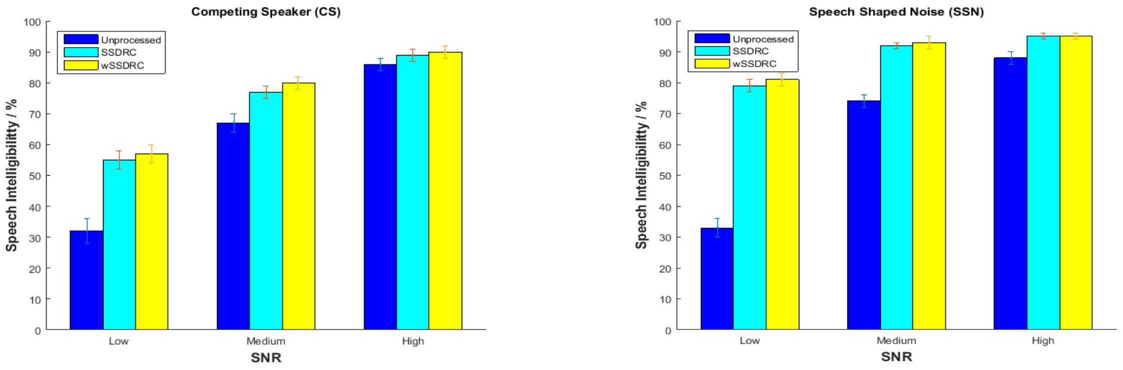
3.1.1 Native English listeners

For this study we recruited N=30 British English native speakers (age range = 18-34; mean age = 25 years). Participants were screened for hearing loss via a Pure Tone Audiometry (PTA), at frequencies of 0.5, 1, 2, 4 kHz. Subjects passed the test with a hearing threshold equal or less than 25 dB HL (averaged across frequencies) in both ears. The listening tests were performed in sound treated booths at University of Edinburgh. Stimuli were presented via Beyerdynamic 770 headphones and participants had to type onto a keyboard what they had heard. For this study SNR levels were set as following: CS Low= -21dB, Mid= -14dB, High= -7dB; SSN Low= -9dB, Mid= -4dB, High= +1dB.

3.1.2 Non-native English listeners

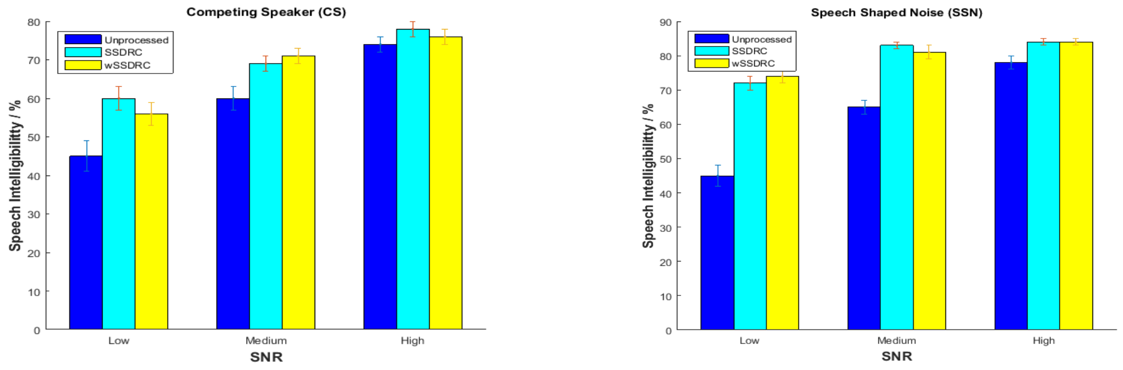
This study was conducted with the participation of students from the University of Crete (UoC). As the subjects were non-native English speakers, they were required to hold a B1 English language qualification. The listening tests were conducted in sound treated booths at the University, and stimuli were administered via high quality

[1] http://speech.ilsp.gr/



(a) (b)

Figure 1. Intelligibility gain for native English listeners



(a) (b)

Figure 2. Intelligibility gains for non-native English listeners

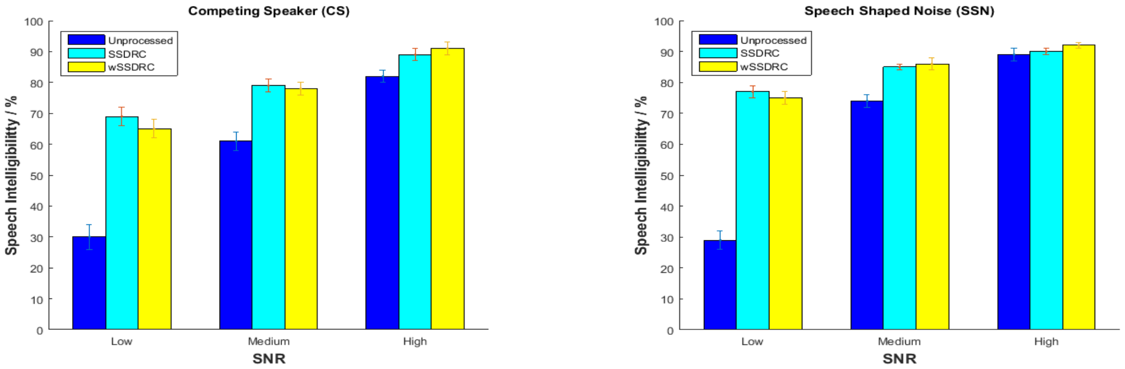
headphones. In total we had N=30 NH participants (average age = 24 years) who were screened for hearing loss in the same way as the native English speakers. In this experiment SNR level were set to: CS Low= -13dB, Mid= -6dB, High= +1dB; SSN Low= -4dB, Mid= +1dB, High= +6dB.

3.1.3 Native Greek NH listeners

For this study, N=21 native Greek NH subjects (average age = 35 years) were recruited to Chania General Hospital, Crete. The participants were screened with a PTA using the same criteria as for the other NH groups. The SNR levels were chosen as following: CS Low= -14dB, Mid= -7dB, High= 0dB; SSN Low= -7dB, Mid= -2dB, High= +3dB.

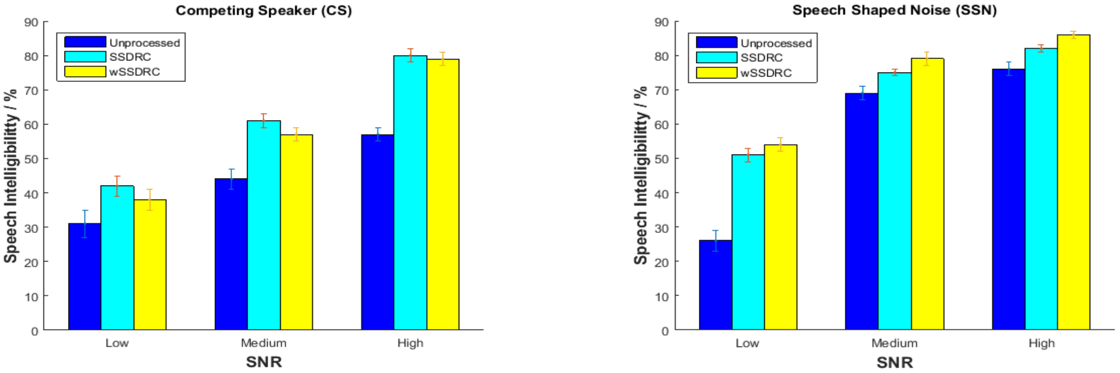
3.1.4 Native Greek HI listeners

N=26 native Greek hearing impaired subjects were recruited at the Chania General Hospital, Crete. Participants were contacted through the Hospital patient database. All of the subjects were screened with a PTA in the range of 0.5-4 kHz, in both ears. The group was characterised by an average hearing loss of 49.8 dB HL (averaged in both ears). Most of the participants were hearing aid wearers and removed their devices while performing the



(a) (b)

Figure 3. Intelligibility gains for normal hearing listeners



(a) (b)

Figure 4. Intelligibility gain for hearing impaired listeners

tests. The age range was 35-60 years, with an average of 49 years. The SNR levels were chosen as follows: CS Low= -9dB, Mid= -6dB, High= 0dB; SSN Low= -4dB, Mid= +1dB, High= +6dB.

Table 1. Speech Quality evaluation as Comparative Mean Opinion Score (CMOS)

Methods	Native Eng.	Non-Native Eng.	Greek NH	Greek HI
Plain	0	0	0	0
SSDRC	-1.6	-1.55	-1.46	-1.15
wSSDRC	-1.64	-1.60	-1.49	-1.24

4 DISCUSSION

Averaged responses from participants in each listening group are plotted in Figure 1-4. Data for SSN and CS noise are plotted separately in each category. As can be seen in Figure 1 and 2, modified speech boosted the intelligibility for both native and non-native listeners, in comparison to unprocessed plain speech (for both

SSDRC and wSSDRC). This change is more noticeable at lower SNR levels, which highlights the relevance of speech modifications in adverse listening scenarios. Larger gains were obtained against the SSN masker. Across all conditions, the wSSDRC model achieved an intelligibility gain equivalent to SSDRC. CMOS scores reported in Table 1 also show a similar trend.

Though both NH and HI listeners benefited from the modifications, the HI group showed larger gains at higher SNR levels (Figure 3 and Figure 4). The same pattern can be seen in CMOS scores in Table 1, where the HI group scored on average 20% higher compared to NH listeners for modified speech. This result might be explained by the reduced sensitivity of HI listeners in higher frequency regions, which might make the unnaturalness of modified speech negligible - if not even more natural-sounding.

5 SUMMARY

In this study we evaluated our recently proposed WaveNet-based intelligibility enhancement (wSSDRC). This neural-based model has the potential to be employed in real-life speech modification applications, which is currently an area of interest for both academia and industry. We used two metrics for our evaluations: CWR for intelligibility and CMOS for quality. This study considered a wide range of listening groups: native or non-native, normal hearing and hearing impaired. Results show that the proposed wSSDRC model considerably improved intelligibility in comparison to plain speech, performing equally well as the traditional SSDRC method, across different listening groups. These findings motivate the idea of extending wSSDRC to real-life speech enhancement applications in future research.

ACKNOWLEDGEMENTS

This work was partly funded by the E.U. Horizon2020 Grant Agreement 675324, Marie Sklodowska-Curie Innovative Training Network, ENRICH.

REFERENCES

[1] Zorila, Tudor-Catalin; Varvara Kandia; Yannis Stylianou. "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression." Thirteenth Annual Conference of the International Speech Communication Association. 2012

[2] Schepker, H.: Rennies, J.; Doclo, S. Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index. The Journal of the Acoustical Society of America, 138(5), 2692-2706.

[3] Simantiraki, Olympia, Martin Cooke, and Simon King. "Impact of Different Speech Types on Listening Effort." Proc. Interspeech 2018 (2018): 2267-2271.

[4] Muhammed Shifas, P. V.; Vassilis Tsiaras; and Yannis Stylianou. "Speech Intelligibility Enhancement Based on a Non-causal Wavenet-like Model." Proc. Interspeech 2018 (2018): 1868-1872.

[5] Cooke, Martin, Catherine Mayo, and Cassia Valentini-Botinhao. "Intelligibility-enhancing speech modifications: the hurricane challenge." Interspeech. 2013.