Improving face-to-face communication in loud environments by means of blind source separation

Martin Berdau[1], Andreas Volgenandt[1], Rainer Huber[1],
Christian Rollwage[1], Jörg Bitzer[1,2], Jan Rennies[1]

[1] *Fraunhofer IDMT, Hearing, Speech and Audio Technology and Cluster of Excellence Hearing4all, Oldenburg, Germany*

[2] *Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, and Cluster of Excellence Hearing4all, Oldenburg, Germany*

Abstract

In this study a Deep Neural Network-based algorithm was used in a hearing protection device with the intention to improve speech intelligibility in noisy environments. The employed long short-term memory (LSTM) network predicts masks in the frequency domain for speech as well as background noise. In order to keep situational awareness, speech was not isolated completely, but rather mixed back with the separated noise at a better signal-to-noise ratio (SNR). Offline evaluations with speech quality and listening effort measures using noises typical for industrial work environments yielded promising results. The algorithm was implemented running in real-time in an electronic earpiece functioning as a hearing protection device. Speech and noise were played back simultaneously and recorded by microphones of the device using the algorithm in order to evaluate the signals afterwards. Results indicate that, while the LSTM approach can potentially improve speech communication in noisy environment, care has to be taken regarding processing delay in relation to the direct sound in order to avoid disturbing echoes.

Motivation

Usually the use of hearing protection devices (HPDs) in noisy environments is required or even regulated by law in order to protect workers, e.g. in the industry, from noise-induced hearing loss. However, several factors exist which reduce the overall acceptance for HPDs. While discomfort caused by the device is an often expressed reason against wearing protection, another huge factor for rejecting the use of HPDs is restricted communication capabilities since fellow workers cannot be understood sufficiently well [2].

Therefore, an algorithm is introduced which aims at increasing speech intelligibility while wearing HPDs. This algorithm is based on blind source separation, which separates speech and noise and allows for mixing those signals back at a higher SNR, thus likely increasing the overall speech intelligibility. The chosen algorithm to perform blind source separation is based on Deep Neural Networks (DNN) [1]. The employed algorithm had previously been used successfully in TV broadcasting to increase speech intelligibility.

Algorithm Description

Figure 1 shows a schematic of the proposed algorithm. The algorithm runs with a sampling frequency of 16 kHz and processes blocks of audio containing 512 samples.

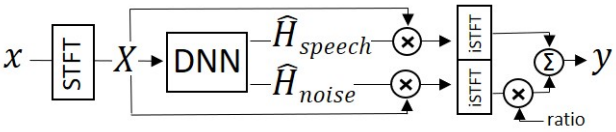

Figure 1: Block diagram of the proposed algorithm.

Short-time Fourier transforms (STFT) are performed to retrieve the input spectrum, which is processed by three subsequent long short-term memory (LSTM) networks, which generate masks for speech and noise. The respective spectra for the estimated speech and noise are calculated by multiplying the masks with the input spectrum. The spectra are transformed back to time-domain and remixed at a desired SNR by attenuating the retrieved noise signal by a fixed ratio.

Instrumental Evaluation

As a first performed analysis signals were processed on a computer, to confirm the applicability of the algorithm in the context of industrial noise environments. Sentences of the Oldenburg Sentence Test (OLSA, [4]) were mixed with industry related noise at SNRs ranging from -10 to +10 dB. These signals were processed by the algorithm and remixed using a background noise attenuation between 2 and 10 dB. Evaluations were made based on the non-intrusive listening effort prediction model presented in [3]. Results are shown in Figure 2. It was found that even small improvements in SNR leads to decreased instrumentally assessed listening effort compared to the unprocessed reference condition. In order to achieve the most effective reduction of listening effort in environments, where the SNR is below 0 dB and listening effort expected to be high, the algorithm was used with an SNR enhancement of 10 dB for further usage, since this approach performed the best among the ones tested for the chosen condition.

Listening Test

In order to evaluate a real-time implementation of the algorithm in combination with an in-ear HPD, a hardware prototype earpiece was used [5]. The device contains two external microphones as well as an internal microphone (facing the ear drum). Sound from the environment can be picked up, processed and played back into the ear canal by integrated drivers. Besides the proposed source separation algorithm, a compressor providing active hearing protection and a filter, that equalizes

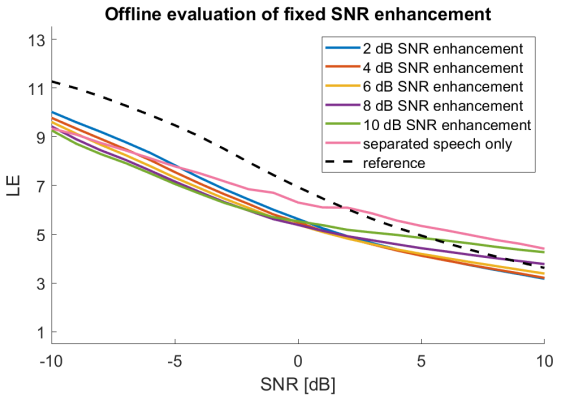

Figure 2: Evaluated listening effort on a computer for several remixes after using the algorithm. For analysis, a model has been used, which computes an Effort Scaling Categorical Unit rating (ESCU) ranging from 1 (no effort) to 13 (extreme effort) [3]. The dotted line indicates the ESCU for the unprocessed input signals of the algorithm.

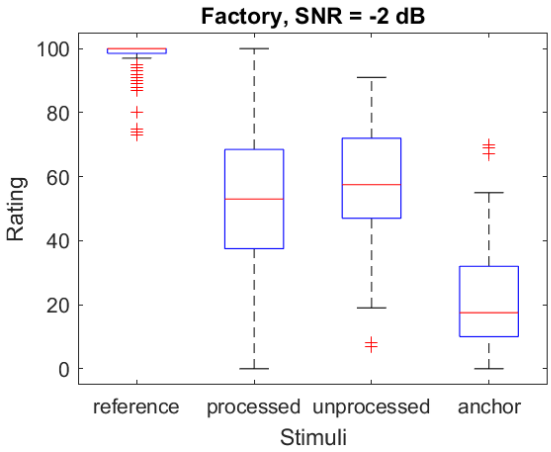

Figure 3: Ratings for factory noise with a noise level of 70 dB(A) and speech level of 68 dB(A) averaged across all four sentences.

transmission properties of the device for a natural hear-through, were used. The earpiece was fitted into a person's ear.

As in the offline analysis, mixtures of speech and noise were used. The mixture was played back using a loudspeaker located 2 m in front of the wearer. In total, three different noises were used, including a rather stationary factory noise as well as an instationary metal grinder and saw noise. Noise was played back at levels of 70, 75, 80 and 85 dB(A) measured at the position of the participant's head. As speech stimuli, four different sentences of four different speakers proposed for conducting audio evaluations were used [6]. Speech signals were played back at levels of 68, 75 and 82 dB(A). Those levels as well as the underlying speech spectra for those levels to simulate the Lombard effect were taken from the ANSI standard describing the Speech Intelligibility Index [7].

The audio was recorded with an outside microphone of the HPD (representing the unprocessed stimuli arriving at the listener's position) as well as with the inside microphone, which contained the stimuli after the entire processing chain (active hearing protection, source separation, and remix with background noise attenuation). In contrast to the offline analysis conducted before, these recordings also contained any leaked direct sound and its possible interaction with the processed sound (e.g., unwanted echo effects due to latency caused by the processing chain).

For evaluating the recorded signals, a method derived from a MUSHRA [8] procedure was used, where participants were instructed to compare four different signals per trial in terms of listening effort on a scale ranging from 0 to 100 units. A rating of 0 indicated the highest possible perceived effort, while a rating of 100 indicated the lowest possible perceived effort. The four test signals included a reference signal, an anchor signal, the unprocessed signal (picked up by the outer microphone), and the processed signal (picked up by the inner microphone). The reference signal was generated by mixing the speech signals with the noise signals at a fixed SNR of +10 dB, while for the anchor signal an 8th order lowpass-filtered version of the unprocessed signal with a cutoff frequency of 1.5 kHz was used. Since the SNR of the reference signal was the highest amongst all signals, this one was expected to be rated highest on the proposed scale. On the other hand, the anchor signal was expected to be rated lowest because of the lowpass. These signals were included to give the participants an orientation when rating the processed and unprocessed signals. Ideally, participants would therefore choose a rating for the processed and unprocessed signals somewhere in-between the rating of reference and anchor signals.

Since a huge amount of audio material was recorded, a subset of conditions was chosen for the evaluation, including the factory noise at SNRs of -2 and -7 dB, as well as the metal grinder noise at an SNR of -17 dB and the saw noise at an SNR of -5 dB. In total, 19 normal-hearing participants aged 21 to 31 years took part in this experiment.

The results for the factory noise with an SNR of -2 dB are shown in Figure 3. As expected, the reference signal got the highest median ratings, while the anchor signal got rated the lowest. When comparing the processed and unprocessed signals, they seemed to induce a similar listening effort for this condition. The interindividual variability, as indicated by the boxes and whiskers, was considerable for these conditions. The corresponding results for the metal grinder noise at an SNR of -17 dB yielded different outcomes (Figure 4). While the reference signal was also rated highest in terms of low listening effort, the unprocessed signal seemed to induce a considerably higher listening effort than the processed signal. The rating for the unprocessed signal was even lower than the rating for the anchor signal, possibly caused by removing high frequency noise in the unprocessed signal with the lowpass-filter. A similar outcome for the saw noise can be seen in Figure 5.

Discussion and Outlook

The proposed algorithm produced no obvious drawback for relatively high SNRs when integrated into a prototyp-

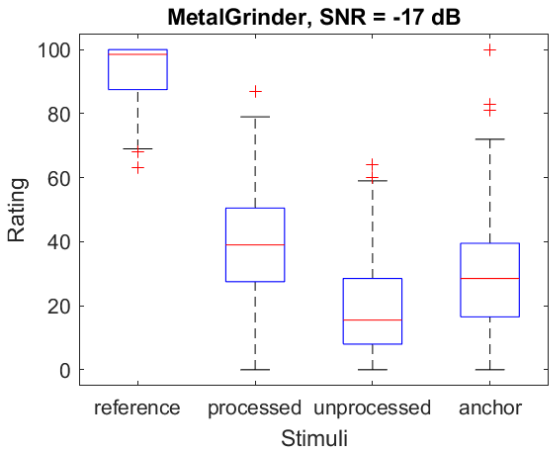

Figure 4: Ratings for metal grinder noise with a noise level of 85 dB(A) and speech level of 68 dB(A) averaged across all four sentences.

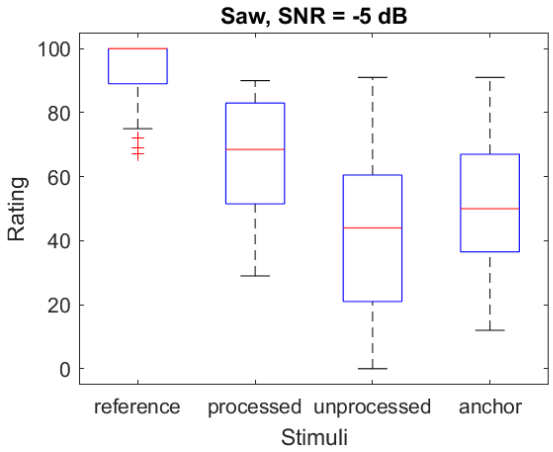

Figure 5: Ratings for saw noise with a noise level of 80 dB(A) and speech level of 75 dB(A) averaged across all four sentences.

ical HPD. In such conditions, intelligibility is expected to be high and listening effort to be relatively low already in the unprocessed condition. In contrast, at very low SNRs, which are not unlikely to occur in industrial environments, the algorithm seemed to be advantageous and clearly reduced the perceived listening effort. However, further experiments might be required to analyze the benefit in more conditions and at a broader range of SNRs.

Using the algorithm in environments with SNRs above 0 dB could lead to an increase in listening effort, as indicated by the offline analysis. This is likely the case because, at higher SNRs, processing artefacts introduced by the source separation have a higher impact on the integrity of the speech signal. It remains to be tested if human listeners are equally sensitive to such artefacts as the instrumental listening analyses measure used in this study.

Another factor that needs further attentions is the processing time of the proposed algorithm. In the present implementation, the delay is about 32 ms, i.e., the direct (leaked) sound and the processed signal can potentially overlap and be perceived as disturbing. In practice, this highly depends on the passive damping properties of the HPD. Although it was not explicitly measured in this study, the listeners did not report disturbing echo effects, so we can assume that processing delay only played a minor role in this study. In practice, such effects (as well as possible artefact-related issues) could possibly be moderated by introducing an adaptive procedure, which activates and deactivates or scales the degree of processing depending on factors like, e.g., SNR or a real-time estimation of perceived listening effort [3].

References

[1] Nils L. Westhausen u. a. „Real-Time Speech Enhancement of TV Broadcast Material by Audio Source Separation based on Recurrent Neural Networks“. In: Transactions on Audio, Speech and Language Processing (under revision).

[2] Rainer Weiß. Beurteilung der Wirksamkeit von Gehörschutz - Befragung von ca. 900 Versicherten aus 37 durch die Süddeutsche Metall-Berufsgenossenschaft betreuten Betrieben. SMBG, 2003.

[3] Huber, Rainer, Melanie Krüger, and Bernd T. Meyer. "Single-ended prediction of listening effort using deep neural networks." Hearing research 359 (2018): 40-49.

[4] Wagener, K., V. Kühnel, and B. Kollmeier. "Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test." Zeitschrift Fur Audiologie 38 (1999): 4-15.

[5] Denk, Florian, et al. "A one-size-fits-all earpiece with multiple microphones and drivers for hearing device research." Audio Engineering Society Conference: 2019 AES INTERNATIONAL CONFERENCE ON HEADPHONE TECHNOLOGY. Audio Engineering Society, 2019.

[6] ITU. P.501 : Test signals for use in telephony and other speech-based applications. URL: `https://www.itu.int/rec/T-REC-P.501/en`.

[7] ANSI S3.5 1997: Methods for Calculation of the Speech Intelligibility Index.

[8] Series, B. "Method for the subjective assessment of intermediate quality level of audio systems." International Telecommunication Union Radiocommunication Assembly (2014).