A Two-Stage End-to-End System for Speech-in-Noise Hearing Aid Processing

Zehai Tu, Jisi Zhang, Ning Ma, Jon Barker

University of Sheffield, Department of Computer Science, Sheffield, UK

{ztu3, jzhang132, n.ma, j.p.barker}@sheffield.ac.uk

Abstract

A two-stage end-to-end system is proposed for hearing aid processing of speech in noise for the Clarity enhancement challenge. In the first stage, a denoising module is optimised for interferer suppression. In the second stage, an amplification module, which is individualised to a listener's hearing ability, is optimised to maximise the intelligibility. The system is built causally, and the latency is lower than 5 ms. The early objective evaluation shows the advantage of the proposed system.

Index Terms: Hearing aid speech processing, speech-in-noise, end-to-end

1. Introduction

The Clarity enhancement challenge [1] aims to find optimal machine learning methods for hearing aid processing of speech-in-noise. This work proposes an end-to-end system, where a denoising module and an amplification module are optimised in two stages. With a differentiable hearing loss model and an intelligibility objective embedded in the optimisation, the system manages to be customised to a listener's hearing ability and improve the intelligibility.

Noise suppression has been used in hearing aids since the 1970s, including adaptive filtering, spectral subtraction, spatial filtering [2, 3, 4]. Many recent hearing aids also include environmental classification algorithms [5, 6] that allow the characteristics of the noise suppression algorithms to be tuned separately for different noise types [3]. Recently, deep neural networks have also achieved impressive success [7, 8, 9].

Hearing aid amplification formulae have long been studied. The National Acoustic Laboratories' Revised (NAL-R) fitting [10] was a well-recognised linear amplification formula. With the introduce of dynamic range compression, more compressive fittings capable are developed, including NAL-NL1, NAL-NL2, CAMEQ, CAMEQ2-HF, DSL [11, 12, 13, 14].

Our recent works [15, 16] have shown the potential of data-driven optimised fittings based on objective evaluations. This work follows the same path and takes advantage of end-to-end learning to optimise a hearing aid processing system for speech in noise.

2. Method

The overall workflow of the method is shown in Fig. 1. For each ear of each hearing impaired listener, a denoising module and an amplification module are optimised to enhance noisy signals in two stages. In the first stage, the denoising module is optimised with a signal-to-noise ratio (SNR) loss. In the second stage, a differentiable hearing loss model is incorporated and the amplification module is optimised with an objective function consisting of an STOI loss [17] and a loudness loss [18]. The denoising module can be optimised jointly in the second stage. All components are implemented with PyTorch [19], and the back-propagation algorithm is used for the optimisation.

2.1. Denoising module

The denoising module D is optimised to suppress the noise and reverberation. Conv-TasNet [7] is an end-to-end convolutional time domain audio separation network and has shown its successes for speech separation and denoising tasks. A Conv-TasNet based multi-channel speech separation approach [20] is used as the denoising module, which has achieved a success for a joint denoising, dereverberation, and separation task. The multi-channel (MC) Conv-TasNet incorporates a spectral encoder, a spatial encoder, a separator and a decoder. A 1-D convolutional layer and a 2-D convolutional layer is used to construct the spectral encoder and the spatial encoder, respectively. Causal convolution and cumulative layer normalisation are used for potential real-time processing. Given a multi-channel noisy signal $x \in \mathbb{R}^{C\times T}$, where C is the number of channels and T is the number of signal samples, the denoising module D estimates the denoised single-channel signal $\hat{y} \in \mathbb{R}^T$.

Different from [7, 20], SNR rather than scale-invariant SNR (SI-SNR) is used as the objective, so that the signal level stays consistent as it is critial for the down-streaming amplification. The SNR loss is expressed as:

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(y,\hat{y}) &= -10\log_{10}\frac{\|y\|^2}{\|y-\hat{y}\|^2+\tau\|y\|^2} \\ &= 10\log_{10}\left(\|y-\hat{y}\|^2+\tau\|y\|^2\right)-10\log_{10}\|y\|^2,\end{aligned} \tag{1}$$

where $\hat{y}$ and y are the estimated and reference signals, respectively, and $\tau = 10^{-\mathrm{SNR}_{\max}/10}$ is a soft threshold preventing examples that are well denoised dominating the gradients within a training batch [21]. $\mathrm{SNR}_{\max}$ is set to 30 dB according to [21].

2.2. Amplification module

The amplification module A aims to implement individualised enhancement to the denoised signals to maximise the intelligibility for the hearing impaired listeners. In this work, both a Conv-TasNet and a finite-impulse response (FIR) filter are experimented to be used as the amplification module. The structure of the amplification Conv-TasNet is roughly consistent with the denoising MC-Conv-TasNet. The amplification FIR is the same as the processor in [15]. The amplification module takes the denoised signal $\hat{y} \in \mathbb{R}^T$ as the input and produces the amplified signal $\hat{z} \in \mathbb{R}^T$.

STOI is used in the objective function as the target is to achieve maximal intelligibility. A loudness constraint term is also included, otherwise the signal could be over-amplified as STOI is based on cross correlation regardless of signal level. The objective function is expressed as:

$$\mathcal{L}_{\mathcal{A}}(y,\hat{z}) = -\mathrm{STOI}(y,\mathrm{HLM}(\hat{z}))+\alpha\|\Gamma(y)-\Gamma(\mathrm{HLM}(\hat{z}))\|^2, \tag{2}$$

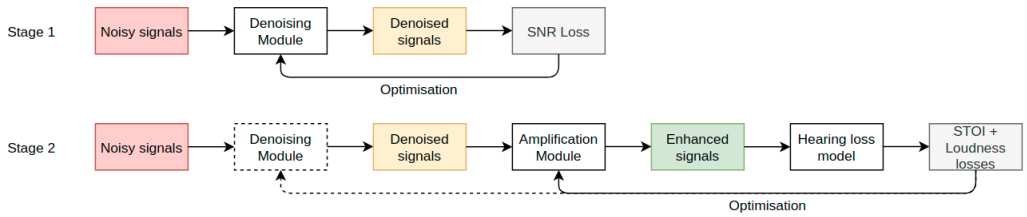

Figure 1: *Overall workflow of the two-stage optimisation for the denoising and the amplification modules*

where α is a weighting coefficient, Γ is the loudness computing formula according to ITU-R BS.1770-4 [22], and HLM represents the hearing loss the model which will be introduced in the next section.

2.3. Hearing loss model

The hearing loss model (HLM) used in this work is a differentiable approximation to the MSBG model [23, 24, 25, 26] released in the challenge, and detailed explained in [16]. Different from the MSBG model, the differentiable hearing loss model takes advantage of FIR filters and Hilbert transformation for fast parallel computing. The model takes the audiogram of the listener as input, and simulates free field, middle- and inner-ear transformation, spectral smearing , and loudness recruitment.

3. Experiments

3.1. Data

The 6000 scenes within the training set are used for the optimisation, and the 2500 scenes within the development set are used for the selection of optimal modules. For each scene, a six-channel signal, which includes the front, mid, and rear microphone inputs for both left and right ear, and a dual-channel clean anechoic signal are provided. The sampling rate of the signals is 44.1 kHz.

For the denoising module optimisation, all six channels are used as the inputs, and one channel of the corresponding anechoic signal is used as the reference. Two denosing modules are optimised separately for the left and the right ear using the left anechoic signal and the right anechoic signal as the references. For the amplification module optimisation, the signal channel output from the denoising module is used as the input, and the corresponding anechoic channel is used the as the reference.

3.2. Setup

A NVIDIA Tesla V100 SXM2 GPU is used for training and inference for each ear of a listener. The signals are downsampled to 22.05 kHz in the optimisation for faster training and inference. In the first optimisation stage, the filter length of the denoising MC-Conv-TasNet is 20, therefore the latency is shorter than 1 ms. The kernel sizes for the spectral and spatial encoder are 256 and 128, respectively. The separator contains six convolutional blocks. Other parameters are the same as the causal configuration in [7].

In the second optimisation stage, both Conv-TasNet and FIR filter are optimised as the amplification module. For Conv-TasNet amplifier, the settings are consistent with the denoising Conv-TasNet, except for the number of separator convolutional blocks being two. For FIR amplifier, the implementation is the same as [15]. The length of the FIR filter is 882, which equivalents to the 4 ms latency. The amplified signals are hard clipped from -1 to 1 after amplification, and then upsampled to 44.1 kHz for the operation of the hearing loss model.

Table 1: *Evaluation results. Amplifier: the amplification module; Joint Opt: whether to optimise the denosing module jointly when optimising the amplification module.*

Amplifier	Joint Opt	MBSTOI	DBSTOI
Baseline	-	0.414	-
OpenMHA	-	0.545	0.650
Conv-TasNet	True	0.645	0.836
Conv-TasNet	False	0.651	0.827
FIR	False	0.646	0.766

3.3. Evaluation

The initial evaluation was conducted on the first listener, i.e. L0001, and the scenes are selected according to the development scenes-listeners list. Both average MBSTOI [27] and DBSTOI [28] scores are computed. MBSTOI is a modified version of DBSTOI to eliminate the predicting offset in low SNRs, but it is based on the assumption of linear and relatively simple scenarios. It is observed that MBSTOI could be invalid in our case, thus DBSTOI is also used for the objective evaluation. The amplification formula in the OpenMHA [29] is also used as the amplification module for comparison. The baseline MBSTOI scores provided by Clarity are also included.

4. Results

The results are shown in Table 1. Conv-TasNet as the amplification module is optimised with and without the joint optimisation of the denoising module. While the FIR will hardly learn anything when jointly optimised with the denoising module, thus only the result of the FIR without joint optimisation is shown. It is observed that joint optimisation with the denoising module will not sufficiently suppress interfering noise, thought it could achieve the highest DBSTOI score.

Conv-TasNet as the amplification module can achieve better objective performance, while it brings more artifacts and corruption to the signal, thus it is submitted for the first stage objective evaluation. On the contrary, FIR as the amplification module achieves slightly lower objective scores, but the enhanced signals are more intelligible according to our initial listening evaluation. Therefore, the FIR enhanced signals are submitted to the second stage subjective evaluation.

5. Conclusions

A two-stage end-to-end system, consisting of a denoising module and an amplification module, is proposed in this work. The early experiments show the intelligibility improvement. Two entries are submitted to the evaluation. The system with Conv-TasNet amplifier can achieve higher objective scores, therefore is submitted to objective evaluation. And the system with FIR

amplifier produces signals with less distortion, therefore is submitted for the final subjective evaluation.

6. References

[1] S. Graetzer, M. Akeroyd, J. P. Barker, T. J. Cox, J. F. Culling, G. Naylor, E. Porter, and R. V. Munoz, "Clarity: Machine learning challenges to revolutionise hearing device processing," *Interspeech*, 2021.

[2] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

[3] R. Bentler and L.-K. Chiou, "Digital noise reduction: An overview," *Trends in Amplification*, vol. 10, no. 2, pp. 67–82, 2006.

[4] H. Levitt, "Noise reduction in hearing aids: A review," *Journal of Rehabilitation Research and Development*, vol. 38, no. 1, pp. 111–122, 2001.

[5] L. Lamarche, C. Giguère, W. Gueaieb, T. Aboulnasr, and H. Othman, "Adaptive environment classification system for hearing aids," *The Journal of the Acoustical Society of America*, vol. 127, no. 5, pp. 3124–3135, 2010.

[6] P. Nordqvist and A. Leijon, "An efficient robust sound classification algorithm for hearing aids," *The Journal of the Acoustical Society of America*, vol. 115, no. 6, pp. 3033–3041, 2004.

[7] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[8] Y. Xia, S. Braun, C. K. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 871–875.

[9] N. L. Westhausen and B. T. Meyer, "Dual-signal transformation lstm network for real-time noise suppression," *arXiv preprint arXiv:2005.07551*, 2020.

[10] D. Byrne and H. Dillon, "The national acoustic laboratories'(NAL) new procedure for selecting the gain and frequency response of a hearing aid," *Ear and Hearing*, vol. 7, no. 4, pp. 257–265, 1986.

[11] G. Keidser, H. Dillon, M. Flax, T. Ching, and S. Brewer, "The nal-nl2 prescription procedure," *Audiology Research*, vol. 1, no. 1, pp. 88–90, 2011.

[12] B. Moore, B. Glasberg, and M. Stone, "Use of a loudness model for hearing aid fitting: III. a general method for deriving initial fittings for hearing aids with multi-channel compression," *British Journal of Audiology*, vol. 33, no. 4, pp. 241–258, 1999.

[13] B. C. Moore, B. R. Glasberg, and M. A. Stone, "Development of a new method for deriving initial fittings for hearing aids with multi-channel compression: CAMEQ2-HF," *International Journal of Audiology*, vol. 49, no. 3, pp. 216–227, 2010.

[14] S. Scollie, R. Seewald, L. Cornelisse, S. Moodie, M. Bagatto, D. Laurnagaray, S. Beaulac, and J. Pumford, "The desired sensation level multistage input/output algorithm," *Trends in Amplification*, vol. 9, no. 4, pp. 159–197, 2005.

[15] Z. Tu, N. Ma, and J. Barker, "DHASP: Differentiable hearing aid speech processing," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 296–300.

[16] ——, "Optimising hearing aid fittings for speech in noise with a differentiable hearing loss model," *arXiv preprint arXiv:2106.04639*, 2021.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[18] C. J. Steinmetz and J. Reiss, "pyloudnorm: A simple yet flexible loudness meter in python," in *Audio Engineering Society Convention 150*. Audio Engineering Society, 2021.

[19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, pp. 8026–8037, 2019.

[20] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On end-to-end multi-channel time domain speech separation in reverberant environments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6389–6393.

[21] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 3846–3857.

[22] B. Series, "Algorithms to measure audio programme loudness and true-peak audio level," 2011.

[23] T. Baer and B. C. Moore, "Effects of spectral smearing on the intelligibility of sentences in noise," *The Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1229–1241, 1993.

[24] ——, "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech," *The Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2277–2280, 1994.

[25] B. C. Moore and B. R. Glasberg, "Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech," *The Journal of the Acoustical Society of America*, vol. 94, no. 4, pp. 2050–2062, 1993.

[26] M. A. Stone and B. C. Moore, "Tolerable hearing aid delays. i. estimation of limits imposed by the auditory path alone using simulated hearing losses," *Ear and Hearing*, vol. 20, no. 3, pp. 182–192, 1999.

[27] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *Speech Communication*, vol. 102, pp. 1–13, 2018.

[28] ——, "Predicting the intelligibility of noisy and nonlinearly processed binaural speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 1908–1920, 2016.

[29] H. Kayser, T. Herzke, P. Maanen, M. Zimmermann, G. Grimm, and V. Hohmann, "Open community platform for hearing aid algorithm research: open master hearing aid (openmha)," *arXiv preprint arXiv:2103.02313*, 2021.