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

This paper summarises an end-to-end system for the first round Clarity enhancement challenge. The system consists of a denoising module and an amplification module for speech-in-noise enhancement for hearing impaired listeners. These two modules are optimised in two stages. In the first stage, the denoising module is optimised for interferer suppression. In the second stage, the amplification module, which is individualised to a listener's hearing ability, is optimised to maximise the intelligibility. Both objective and subjection evaluation results show the system with the configuration incorperating a multi-channel Conv-TasNet based denoising module and a finite impulse response filter based amplification module can significantly outperform the baseline system.

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

1. Introduction

The Clarity challenge [1] aims to find optimal machine learning methods for hearing aid processing of speech-in-noise. In the first round of enhancement challenge, the problem of speech-in-noise in everyday home environments is addressed. Domestic scenes containing hearing impaired listeners, target speech signals and noise or speech interferers are simulated. Participating systems are expected to provide enhancement for the listeners and maximise intelligibility of the target speech signals.

This paper describes the Sheffield system, which consists of a denoising module and an amplification module, targeting at interferer suppression and hearing loss compensation. 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.

This paper is organised as follows. Section 2 briefly reviews related work on noise suppression for hearing aids and hearing aid amplification formulae. Section 3 describes the proposed two-stage end-to-end enhancement system. Section 4 presents the database, the detailed system setup, and the evaluation methods. The results comparing the performances of proposed systems and the baseline system are presented and discussed in Section 5. Section 6 concludes the paper.

2. Background

Noise suppression has been used in hearing aids since the 1970s, including adaptive filtering, spectral subtraction, spatial filtering [2, 3, 4], in which beamforming is particularly popular [5]. Many recent hearing aids also include environmental classification algorithms [6, 7] 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 [8, 9, 10].

Hearing aid amplification formulae have long been studied. The National Acoustic Laboratories' Revised (NAL-R) fitting [11] 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 [12, 13, 14, 15].

Our recent works [16, 17] 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.

3. Method

The overall workflow of the method is shown in Fig. 1. For each ear of a hearing impaired listener, a denoising module $\mathcal{M}_\mathcal{D}$ and an amplification module $\mathcal{M}_\mathcal{A}$ are optimised to enhance noisy signals. The overall optimisation is divided into two stages. In the first stage, $\mathcal{M}_\mathcal{D}$ is optimised with a signal-to-noise ratio (SNR) loss for noise and reverberation suppression. In the second stage, a differentiable hearing loss model $\mathcal{M}_\mathcal{HL}$ is incorporated, and $\mathcal{M}_\mathcal{A}$ is optimised with an objective function consisting of an STOI loss [18] and a loudness loss [19] for the compensation of hearing impairment. Meanwhile, $\mathcal{M}_\mathcal{D}$ can be jointly optimised in the second stage. All components are implemented with PyTorch [20], and the back-propagation algorithm is used to compute gradients for the optimisation. $\mathcal{M}_\mathcal{D}$, $\mathcal{M}_\mathcal{A}$ and $\mathcal{M}_\mathcal{HL}$ are described in this section.

3.1. Denoising module

The denoising module $\mathcal{M}_\mathcal{D}$ aims to suppress disturbances caused by both noise and speech interferers. Conv-TasNet [8] is an end-to-end convolutional time domain audio separation network and has shown its successes for single-channel speech separation and denoising tasks. In order to exploit the spatial information provided by multi-channel signals in the Clarity Challenge, the multi-channel (MC) Conv-TasNet is used in this work as $\mathcal{M}_\mathcal{D}$. The MC-Conv-TasNet has been proved effective for a joint denoising, dereverberation and separation task [21].

The structure of MC-Conv-TasNet $\mathcal{M}_\mathcal{D}$ is shown in Fig 2. It incorporates a spectral encoder, a spatial encoder, a separator and a decoder. 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 spectral encoder takes one channel as the input and maps segments of this channel $x_0 \in \mathbb{R}^{1\times T}$ to high-dimensional features with a 1-D convolutional layer. Meanwhile, the spatial encoder extracts the spatial information from x with a 2-D convolutional layer. Outputs of both spectral and spatial encoders are utilised by the separator, which then computes a mask for the target features. The separator consists of multiple 1-D convolutional blocks, which includes multiple 1-D convolutional layers, PReLU activations, normalisation layers, and residual connections. Finally, the decoder reconstructs a

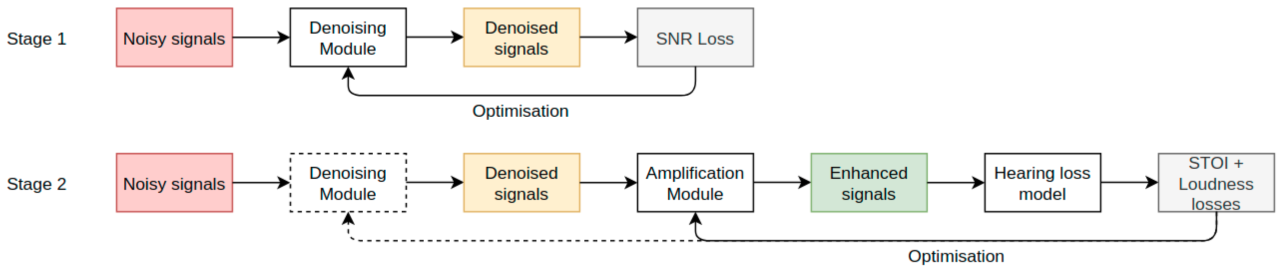

Figure 1: *Overall workflow of the two-stage optimisation for the denoising and the amplification modules.*

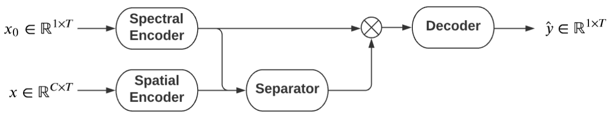

Figure 2: *Structure of MC-Conv-TasNet*

single channel output $\hat{y} \in \mathbb{R}^{1\times T}$ with the estimated features provided by the separator.

Different from [8, 21], SNR rather than scale-invariant SNR (SI-SNR) is used as the objective, so that the signal level stays consistent as it is critical for the down-streaming amplification. The SNR loss is expressed as:

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

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 [22]. $\mathrm{SNR}_{\max}$ is set to 30 dB according to [22].

3.2. Amplification module

The amplification module $\mathcal{M}_{\mathcal{A}}$ 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 [16]. The amplification module takes the denoised signal $\hat{y} \in \mathbb{R}^{1\times T}$ as the input and produces the amplified signal $\hat{z} \in \mathbb{R}^{1\times T}$.

STOI [23] 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:

$$\begin{aligned}\mathcal{L}_{\mathcal{A}}(y,\hat{z}) &= -\mathrm{STOI}(y,\mathcal{M}_{\mathcal{HL}}(\hat{z})) \\ &\quad +\alpha\|\Gamma(y)-\Gamma(\mathcal{M}_{\mathcal{HL}}(\hat{z}))\|^2,\end{aligned} \tag{2}$$

where α is a weighting coefficient, Γ is the loudness computing formula according to ITU-R BS.1770-4 [24], and $\mathcal{M}_{\mathcal{HL}}$ represents the hearing loss the model which will be introduced in the next section.

3.3. Hearing loss model

The hearing loss model $\mathcal{M}_{\mathcal{HL}}$ used in this work is a differentiable approximation to the MSBG model [25, 26, 27, 28] released in the challenge, and detailed explained in [17]. 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 a listener as input, and simulates free field, middle- and inner-ear transformation, spectral smearing , and loudness recruitment.

4. Experimental setup

4.1. Databases

4.1.1. Scene databases

The Clarity challenge provides 10,000 simulated scenes, 6,000 of which are used as training set (*train*), 2,500 are treated as the development set (*dev*), and 1,500 are used for final evaluation set (*eval*). Utterances from 24 speakers are selected for *train*, 10 for *dev*, and 6 for *eval*. Each scene incorporates a six-channel noisy signal, which consists of the front, mid, and rear microphone inputs for both left and right ear, and a dual-channel clean anechoic signal. The sampling rate of the signals is 44.1 kHz. Half of the speech interferers are domestic noises, and the other half are speeches of a second speaker.

4.1.2. Listener databases

Bilateral pure-tone audiograms are used to characterise listeners' hearing abilities by recording the hearing thresholds at [250, 500, 1000, 2000, 3000, 4000, 6000, 8000] Hz. 100 audiograms are provided in *train* and *dev*, and another 50 audiograms for the *eval*.

4.2. System setup

4.2.1. Denoising module

The network configuration of $\mathcal{M}_{\mathcal{D}}$ is described in this section. 256 and 128 filters are used in the spectral and the spatial encoders, respectively. The length of the encoder filters is 20 samples. 256 and 512 channels are used in the bottleneck 1×1 convolutional block and the convolutional blocks, respectively. The kernel size in the convolutional blocks is 3. 6 convolutional blocks with dilation factors of 1, 2, 4, ..., 32 are repeated 4 times within the separator.

All six channels of noisy signals are used as the input, and one channel of anechoic signals is used as the reference (dependent on left or right ear). The signals are downsampled to 22.05 kHz. $\mathcal{M}_{\mathcal{D}}$ is trained for 200 epochs on 2-second long segments. Adam optimiser [29] is used for training with the initial learning rate of $1e$-3. Gradient clipping with maximum L2-norm of 5 is applied. The convolution layers and layer normalisations in $\mathcal{M}_\mathcal{D}$ are implemented causally. A NVIDIA Tesla V100 SXM2 GPU is used for training $\mathcal{M}_\mathcal{D}$, and two modules are trained in total for the left and right ear.

4.2.2. Amplification module

Both Conv-TasNet and FIR filter are optimised as the amplification module, noted as $\mathcal{M}_\mathcal{A}^\mathcal{C}$ and $\mathcal{M}_\mathcal{A}^\mathcal{F}$, respectively. As hearing losses cause sophisticated non-linear degradation, $\mathcal{M}_\mathcal{A}^\mathcal{C}$ is expected to provide such an amplification that can be better fit to this degradation. In contrast, $\mathcal{M}_\mathcal{A}^\mathcal{F}$ is optimised to provide a simple and linear amplification which processes signals with constraints, i.e., avoid distortion or artifacts. The configuration of $\mathcal{M}_\mathcal{A}^\mathcal{C}$ is consistent with $\mathcal{M}_\mathcal{A}$, except for the number of separator convolutional blocks being two. The implementation of $\mathcal{M}_\mathcal{A}^\mathcal{F}$ is detailed described in [17], and the length of the FIR filter is 882. The latency of $\mathcal{M}_\mathcal{A}^\mathcal{C}$ is less than 1 ms as the encoder filter length is 20. The latency of $\mathcal{M}_\mathcal{A}^\mathcal{F}$ used for evaluation is more than 5 ms, while we further reduced the tap size of the FIR filter to 220 and the difference is minimal.

The single-channel output of $\mathcal{M}_\mathcal{D}$ is used as the input, and $\mathcal{M}_\mathcal{A}$ produces a single-channel amplified signal for hearing loss compensation to each ear. The amplified signals are hard clipped from -1 to 1 after amplification, and then upsampled to 44.1 kHz for the processing of $\mathcal{M}_{\mathcal{HL}}$. $\mathcal{M}_\mathcal{A}^\mathcal{C}$ is trained for 50 epochs with the initial learning rate of $1e$-3 and $\mathcal{M}_\mathcal{A}^\mathcal{F}$ is trained for 20 epochs with the learning arate of $5e$-2.

4.3. Evaluation

4.3.1. Preliminary evaluation

The preliminary evaluation was conducted on the first listener in the *dev* set, i.e. L0001, and the scenes are selected according to the development scenes-listeners list. Following the objective evaluation methods conducted by the challenge, enhanced signals are processed by the MSBG hearing loss model and compared with the corresponding anechoic siganls. The differences are measured by both MBSTOI [30] and DBSTOI [31] scores. 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, therefore DBSTOI is also used for the objective evaluation. The amplification formula in the OpenMHA [32] is also used as the amplification module for comparison. The baseline MBSTOI scores provided by Clarity are also included.

4.3.2. Final evaluation

The final evaluation consists of an objective evaluation and a subjective evaluation, both conducted by the Clarity committee. In the objective evaluation, each scene within *eval* is evaluated with four audiograms by MBSTOI. In the subjective evaluation, each scene within *eval* is evaluated by a hearing impaired listener with recognition tests.

5. Results

5.1. Preliminary results

The preliminary results are shown in Table 1. The baseline uses OpenMHA system as $\mathcal{M}_\mathcal{A}$ without noise suppression as shown in the first row. It is clear that baseline system can benefit from

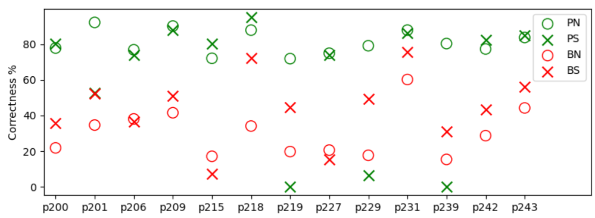

Figure 3: *Subjective evaluation scores for each listener. PN: proposed system on scenes with noise interferers; PS: proposed system on scenes with speech interferers; BN: baseline system on scenes with noise interferers; BS: baseline system on scenes with speech interferers.*

Table 1: *Preliminary evaluation results.* $\mathcal{M}_\mathcal{D}$*: the denoising module;* $\mathcal{M}_\mathcal{A}$*: the amplification module; Joint Opt: whether to optimise the denosing module jointly when optimising the amplification module.*

$\mathcal{M}_\mathcal{D}$	$\mathcal{M}_\mathcal{A}$	Joint Opt	MBSTOI	DBSTOI
-	OpenMHA	-	0.414	-
MC-Conv-TasNet	OpenMHA	-	0.545	0.650
MC-Conv-TasNet	Conv-TasNet	True	0.645	0.836
MC-Conv-TasNet	Conv-TasNet	False	0.651	0.827
MC-Conv-TasNet	FIR	False	0.646	0.766

the noise suppression provided by MC-Conv-TasNet. When $\mathcal{M}_\mathcal{A}^\mathcal{C}$ is used as the amplification module, it can hardly gain benefits from joint optimisation. And $\mathcal{M}_\mathcal{A}^\mathcal{F}$ will hardly learn nothing when jointly optimised with $\mathcal{M}_\mathcal{D}$, thus the results are not shown here. In conclusion, joint optimisation of $\mathcal{M}_\mathcal{D}$ does not bring significant improvement to the overall system.

It can be observed that $\mathcal{M}_\mathcal{A}^\mathcal{C}$ as the amplification module can achieve better objective performance, while it brings more artifacts and corruption to the signal, therefore it is submitted for only the 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 final subjective evaluation.

5.2. Objective results

Table 2: *Final* ***objective*** *results.* $\mathcal{M}_\mathcal{D}$*: MC-Conv-TasNet based denoising module;* $\mathcal{M}_\mathcal{A}^\mathcal{C}$*: Conv-TasNet based amplification module;* $\mathcal{M}_\mathcal{A}^\mathcal{F}$*: FIR based amplification module.*

Method	Speech interferer		Noise interferer		Overall	
	Median	Mean	Median	Mean	Median	Mean
Baseline	0.33	0.34	0.28	0.29	0.31	0.31
$\mathcal{M}_\mathcal{D}$ + $\mathcal{M}_\mathcal{A}^\mathcal{C}$	0.70	0.70	0.67	0.67	0.69	0.69
$\mathcal{M}_\mathcal{D}$ + $\mathcal{M}_\mathcal{A}^\mathcal{F}$	0.74	0.73	0.69	0.69	0.72	0.71

The results of MBSTOI objective evaluation are shown in Table 2. The proposed systems achieve significant improvement over the baseline. Both $\mathcal{M}_\mathcal{A}^\mathcal{C}$ and $\mathcal{M}_\mathcal{A}^\mathcal{F}$ perform better on scenes with speech interferers than with noise interferers. It is worth noting that $\mathcal{M}_\mathcal{A}^\mathcal{F}$ performs better than $\mathcal{M}_\mathcal{A}^\mathcal{C}$, as FIR filter has such a simple structure that could have better generalisation ability, compared to deep neural network based $\mathcal{M}_\mathcal{A}^\mathcal{C}$.

5.3. Subjective results

Table 3: *Final* ***subjective*** *results.* $\mathcal{M}_\mathcal{D}$*: MC-Conv-TasNet based denoising module;* $\mathcal{M}_\mathcal{A}^\mathcal{F}$*: FIR based amplification module.*

Method	Correctness (per cent)		
	Speech interferer	Noise interferer	Overall
Baseline	43.98	30.30	37.13
$\mathcal{M}_\mathcal{D} + \mathcal{M}_\mathcal{A}^\mathcal{F}$	61.88	81.03	71.45

The overall subjective evaluation results are shown in Table 3. The recognition results of each listener on scenes with speech and noise interferers are shown in Figure 3. Overall speaking, the proposed $\mathcal{M}_\mathcal{D} + \mathcal{M}_\mathcal{A}^\mathcal{C}$ system significantly outperform the baseline system. It is worth noting that the proposed system achieves higher recognition correctness on scenes with noise interferers than speech interferers. This can be explained by observing the individual performances. The recognition correctnesses of three listeners (p219, p229, p239) on scenes with speech interferers are exceptionally low, which could be caused by the confusion on the listening test instruction, i.e., the listeners are asked to repeat what the second talker when the interferer is speech, and the system could completely eliminate the interfering speech thus there is only one talker left, which could be the cause of confusion.

6. Conclusions

A two-stage end-to-end system, consisting of a denoising module $\mathcal{M}_\mathcal{D}$ and an amplification module $\mathcal{M}_\mathcal{A}$, is proposed in this work. Both objective and subjective evaluation results show that the combination of MC-Conv-TasNet based denoising module and FIR filter based amplification module can significantly outperform the baseline system.

7. 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] S. Doclo, S. Gannot, M. Moonen, A. Spriet, S. Haykin, and K. R. Liu, "Acoustic beamforming for hearing aid applications," *Handbook on array processing and sensor networks*, pp. 269–302, 2010.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

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

[25] 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.

[26] ——, "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.

[27] 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.

[28] 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.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] 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.

[31] ——, "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.

[32] 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.