# Sheffield System for the Second Clarity Enhancement Challenge

*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 report describes the Sheffield system for the 2nd Clarity Enhancement Challenge (CEC2) concerning maximising speech intelligibility for hearing impaired listeners. The CEC2 database provides a large number of simulated domestic scenes, each of which contains a target speech degraded by two or three interfering sources, e.g. domestic noises, music, and other speech signals. An enhancement system is wanted to produce an enhanced binaural speech signal given three pairs of binaural noisy speech signals with an ideal latency within 5 ms.

The Sheffield system consists of a denoising module and an amplification module. The denoising module aims to suppress interfering sources and restore target speech, and the amplification module is optimised to compensate for hearing losses. We take advantage of a causal multi-channel densely connected convolutional U-Net with target speaker extraction for denoising, with a Short-time Fourier Transform (STFT) window introducing 4 ms latency. For amplification, we use a simple neural network mapping the audiogram of an impaired ear to a finite impulse response (FIR) filter, which introduces less than 4 ms latency. The overall ideal latency of the proposed cascaded system thus meets the requirement.

## 1. Method

The denoising and amplification modules are described in this section. As shown in Fig 1, a two-stage training strategy is used in this work, i.e., the two modules are optimised separately with different loss functions. The denoising module takes a multi-channel noisy scene signal and a target speaker adaption speech as input and produces binaural enhanced speech. The amplification module takes the denoised speech and audiograms of target hearing impaired listeners as input, and outputs the overall enhanced speech. The echoic target signals of the first binaural channel are used as the labels.

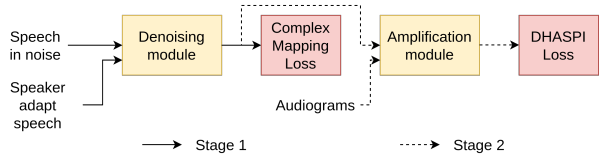


Figure 1: *Overall workflow of the two-stage training for the extraction model and the amplification model.*

### 1.1. Denoising module

In the first Clarity Enhancement Challenge (CEC1) [1], speech enhancement techniques have been demonstrated to benefit hearing aid systems [2, 3]. Specifically, a deep learning based time-domain speech enhancement system with very low-latency plays a crucial role in the winning system [3]. However, time-domain neural enhancement systems usually introduce large artificial distortions to enhanced speech signals. In this work, we propose a frequency domain enhancement approach, extraction DenseUNet (Extr-DenseUNet), which exploits speaker identity information to enhance target speech signals for the hearing aid system.

The Extr-DenseUNet consists of a speaker embedding network to generate speaker identity representation and a speaker extraction network that exploits the speaker identity information to recover target speaker's speech component given a noisy mixture. The speaker embedding network and the extraction network are trained jointly to optimise a signal reconstruction loss, i.e., complex spectral mapping (CSM) proposed in [4]. The speech enhancement system directly outputs a binaural signal.

The speaker embedding network takes as input an enrollment speech signal from a target speaker to generate an embedding vector that represents the speaker identity. Given an enrollment speech signal, STFT is applied to transform the signal to time-frequency (T-F) domain representations, and the magnitude of each T-F bin is used as input features for the speaker embedding network. The speaker embedding network uses a temporal convolutional network (TCN) [5] block to model the input sequential feature. The TCN is built from R repetitions of a sub-block which stacks X dilated 1-D convolutional blocks. The output of the TCN is processed by a standard 1-D convolutional layer followed by a time-averaging operation.

The speaker extraction network consists of two components: a speaker stack and a separation stack. Given a six-channel mixture signal, a STFT with 4 ms window length and 2 ms hop length is used to transform the signal to T-F domain representations with a FFT size of 512. The real and imaginary (RI) components of multi-channel T-F representations are concatenated as input to the extraction network. The design of speaker stack is motivated by a time-domain speaker extraction system [6], which aims to process the speaker representations to coordinate with the main separation stack. The speaker stack employs a TCN block and a 2-D convolutional layer to process the input multi-channel mixture features. Then, the output features from the 2-D convolutional layer are modulated with a target speaker embedding vector through element-wise multiplication to obtain the final speaker information features.

The separation stack uses a dense U-Net consisting of an encoder-decoder structure, which has been successfully developed for both speech enhancement [4] and speaker extraction tasks [7]. Both the encoder and decoder are constructed from four densely connected convolutional blocks. Between the encoder and decoder are two TCN blocks to model long-range temporal information. To received the speaker identity information, the features output from the first dense block in the encoder are concatenated along the channel axis with the speaker infor-

*Equal contribution

Table 1: *Dev set BEHASPI evaluation results.*

| Denoising | | | Amplification | | BEHASPI |
|---|---|---|---|---|---|
| Model | Spk | Loss | NAL-R | Optimised | |
| - | - | - | - | - | 0.1615 |
| - | - | - | ✓ | - | 0.2492 |
| MC-ConvTasNet [3] | - | SNR | ✓ | - | 0.3014 |
| Extr-DenseUNet | ✓ | CSM | ✓ | - | 0.4209 |
| Extr-DenseUNet | ✓ | CSM | - | ✓ | 0.5088 |

mation features from the speaker stack. All the convolution and normalisation layers in the extraction network are causal, and the overall latency of the enhancement system is 4 ms coming from the STFT.

### 1.2. Amplification module

The amplification module targets to amplify a denoised speech signal in a way that help improves the intelligibility. It contains a three layer neural network, which maps an audiogram to six amplification coefficients at [250, 500, 1000, 2000, 4000, 6000] Hz. The amplification coefficients are then converted to a FIR filter that is applied to the denoised signal.

Similarly to our CEC1 system, the amplification module is optimised to maximise the objective evaluation metric, which is the better-ear HASPIv2 [8] (BEHASPI) in CEC2. A differentiable approximation to HASPIv2 is implemented and used as the loss function, together with an energy constraint term to prevent over-amplification, similar to [9]. The DHASPI loss implementation in this work is a refined version from the one in [9]. The differences between DHASPI and HASPIv2 are:

- For the purpose of faster optimisation, the IIR gammatone filterbank is replaced by FIR implementation, and inner-hair cell adaptation and group delay compensation are not included.
- Internal noise are not added as DHASPI is served as an optimisation objective.
- The alignment within HASPIv2 are not included in DHASPI, as it is assumed the denoised signals and reference signals are aligned before amplification.
- The voice activity detection within the feature extraction process is replaced by zero masking.
- The ensemble of neural networks mapping the ten correlation features to the predicted HASPI scores is replaced by another re-optimised neural network.

## 2. Experimental results

The experimental results on the dev set are shown in the Table 1. The first two results show the BEHASPI score of unprocessed speech signals and the NAL-R [10] amplified speech. The denoising Extr-DenseUNet can outperform the MC-ConvTasNet [3] in terms of BEHASPI when the amplification is the NAL-R fitting. The performance can be further improved when using the proposed optimised amplification module.

## 3. References


[1] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, R. Viveros Munoz *et al.*, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in *INTERSPEECH*, 2021.

[2] S. J. Yang, S. Wisdom, C. Gnegy, R. F. Lyon, and S. Savla, "Listening with Googlears: Low-latency neural multiframe beamforming and equalization for hearing aids," in *Proc. Clarity*, 2021.

[3] Z. Tu, J. Zhang, N. Ma, and J. Barker, "A two-stage end-to-end system for speech-in-noise hearing aid processing," in *Proc. Clarity*, 2021.

[4] Z.-Q. W. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1778–1787, 2020.

[5] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*, 2016.

[6] J. Zhang, C. Zorila, R. Doddipatla, and J. Barker, "Time-domain speech extraction with spatial information and multi speaker conditioning mechanism," in *ICASSP*, 2021.

[7] J. Han, Y. Long, L. Burget, and J. H. Cernocký, "DPCCN: Densely-connected pyramid complex convolutional network for robust speech separation and extraction," in *ICASSP*, 2022.

[8] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (haspi) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[9] Z. Tu, N. Ma, and J. Barker, "DHASP: Differentiable hearing aid speech processing," in *ICASSP*, 2021.

[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.