

# MULTI-CHANNEL TARGET SPEAKER EXTRACTION WITH REFINEMENT: THE WAVLAB SUBMISSION TO THE SECOND CLARITY ENHANCEMENT CHALLENGE

Samuele Cornell<sup>1,2</sup>, Zhong-Qiu Wang<sup>2</sup>, Yoshiki Masuyama<sup>3,2</sup>, Shinji Watanabe<sup>2</sup>  
Manuel Pariente<sup>4</sup>, Nobutaka Ono<sup>3</sup>

<sup>1</sup>Università Politecnica delle Marche, Italy <sup>2</sup>Carnegie Mellon University, USA

<sup>3</sup>Tokyo Metropolitan University, Japan <sup>4</sup>Pulse Audition, France

{cornellsamuele, wang.zhongqiu41}@gmail.com

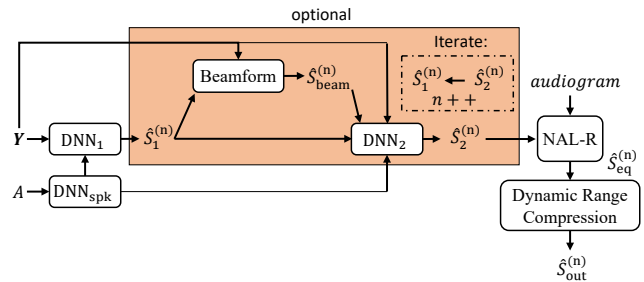
## ABSTRACT

This paper describes our submission to the Second Clarity Enhancement Challenge (CEC2), which consists of target speech enhancement for hearing-aid (HA) devices in noisy-reverberant environments with multiple interferers such as music and competing speakers. Our approach builds upon the powerful iterative neural/beamforming enhancement (iNeuBe) framework introduced in our recent work, and this paper extends it for target speaker extraction. We therefore name the proposed approach as iNeuBe-X, where the “X” stands for extraction. To address the challenges encountered in the CEC2 setting, we introduce four major novelties: (1) we extend the state-of-the-art TF-GridNet model [1], originally designed for monaural speaker separation, for multi-channel, causal speech enhancement, and large improvements are observed by replacing the TCNDenseNet used in iNeuBe [2] with this new architecture; (2) we leverage a recent dual window size approach with future-frame prediction [3] to ensure that iNeuBe-X satisfies the 5 ms constraint on algorithmic latency required by CEC2; (3) we introduce a novel speaker-conditioning branch for TF-GridNet to achieve target speaker extraction; and (4) we propose a fine-tuning step, where we compute an additional loss with respect to the target speaker signal compensated with the listener audiogram. Without using external data, on the official development set our best model reaches a hearing-aid speech perception index (HASPI) score of 0.942 and a scale-invariant signal-to-distortion ratio improvement (SI-SDRi) of 18.8 dB. These results are promising given the fact that the CEC2 data is extremely challenging (e.g., on the development set the mixture SI-SDR is  $-12.3$  dB). A demo of our submitted system is available at WAVLab\_CEC2\_demo.

## 1. PROPOSED METHODS

### 1.1. iNeuBe-X System Overview

Our proposed iNeuBe-X system is depicted in Fig. 1. It is motivated by the iNeuBe framework [2], which contains two multi-microphone input single-microphone output (MISO) DNNs [4], with the first one producing an initial target estimate for beamforming and the second one performing post-filtering for better enhancement. One key difference is that here we condition each MISO DNN with a speaker-enrollment embedding extracted from a third DNN<sub>spk</sub> module, which is jointly trained with the rest and takes as input an enrollment utterance of the target speaker. Given the complex multi-channel short-time Fourier transform (STFT) spectrogram of the mixture signal,  $\mathbf{Y}$ , the first DNN (denoted as DNN<sub>1</sub> in Fig. 1) estimates the complex STFT spectrogram  $\hat{S}_1^{(n)}$  (at iteration  $n = 1$ ) of the anechoic target speech captured at a reference HA array mic, which is set to “CH1 Left” in CEC2. This estimate can then be refined by a second step, which involves a DNN-supported beamformer, and a second MISO DNN (denoted as DNN<sub>2</sub>) with additional inputs from the outputs of



**Fig. 1:** Overview of proposed iNeuBe-X framework. We employ a causal multi-channel Wiener filter beamformer between DNN<sub>1</sub> and DNN<sub>2</sub>.

DNN<sub>1</sub> and the beamformer. This refinement can be iteratively performed, following [2]. Our system is trained to operate with a 32 kHz sampling rate and with an STFT with 16 ms window, 4 ms stride, and the square-root Hann window. This would give such system a theoretical latency of 16 ms [3]. To overcome this, we train both DNN<sub>1</sub> and DNN<sub>2</sub> to predict the current plus future 3 frames of the target anechoic signal, so that a new synthesized frame output can be obtained for each 4 ms stride. This is done in practice by padding zeros to the left of the input mixture STFT. The beamforming operation is performed in a frame-online fashion to comply with the 5 ms constraint on algorithmic latency required in CEC2. We employ a causal multi-channel Wiener filter as in [2, 3]. Since the listener head can rotate in CEC2, instead of simply accumulating the statistics, we use the recursive averaging strategy [5] with a 0.5 forgetting factor.

### 1.2. Speaker Enrollment Embedding Extraction

For DNN<sub>spk</sub>, we use the encoder and TCN modules (and remove the decoder) of the TCNDenseNet architecture [2, 6] to extract speaker embeddings. As the enrollment utterances are available beforehand, this DNN can be non-causal and its output can be cached before inference for each target speaker. Given a complex spectra  $A$  of a monaural adaptation utterance, this DNN extracts a fixed-length speaker embedding  $X_{\text{adapt}} \in \mathbb{R}^{128}$  by performing mean-pooling over time on the outputs of the TCN module. The speaker embedding is then leveraged as an additional cue by the iNeuBe framework for target extraction. Differently from [2, 6], we use a smaller DNN, with 3 TCN repeats, 4 TCN blocks, 128 TCN channels, and 16 hidden channels in the down-sampling Conv2D and DenseNet layers of the encoder. This amounts to a total of around 0.6 million (M) parameters.

### 1.3. Frame-online Speaker-conditioned MISO-TF-GridNet

For both DNN<sub>1</sub> and DNN<sub>2</sub> we use TF-GridNet [1], which is a state-of-the-art separation model that recently achieved an impressive 23.4 dB SI-SDRi on the WSJ0-2mix dataset [7]. Motivated by this strong result, we extended this model to multi-channel scenarios as

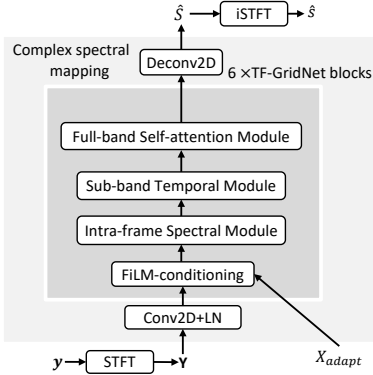


Fig. 2: Proposed frame-online speaker-conditioned MISO-TF-GridNet.

depicted in Fig. 2. Since it uses complex spectral mapping, we feed all channels by simply stacking the real and imaginary (RI) components of all input channels [2] to TF-GridNet, converting it to a MISO model named as MISO-TF-GridNet. We use feature-wise linear modulation (FiLM) [8] over the speaker embedding extracted by  $DNN_{\text{spk}}$  at the beginning of each TF-GridNet block to condition the separation. Note that TF-GridNet proposed in [1] is non-causal. Here we make the following modifications to achieve frame-online processing. Firstly, we use layer normalization (LN) after the first Conv2D layer. Secondly, we use a causal full-band self-attention module by masking the attention matrix. Thirdly, we use a modified sub-band temporal module, where we employ an unidirectional LSTM (instead of bidirectional as in [1]) and, in the unfolding operation, we stack only the previous frames’ features with the current. This way, the sub-band temporal module can output one frame for each frame in the input if the stride is set to 1.

We use the same hyper-parameters suggested in [1], except that  $E$  is set to 2 here. The resulting TF-GridNet has around 8 M parameters and the full iNeube-X model (with  $DNN_1$  and  $DNN_2$ ) roughly doubles that amount.

#### 1.4. STFT-domain NAL-R and Dynamic Range Compression

We found that enhancement alone is not sufficient for obtaining an high HASPI. Without any compensation to listener’s hearing loss, the HASPI for the oracle target speech is less than 0.7 on the development set. To perform fitting in a frame-online fashion, we perform the NAL-R fitting [9] and dynamic range compression in the STFT domain. We followed the dynamic range compression algorithm implemented in [10], and set the knee to -10 dB, threshold -40 dB, ratio 1.2, knee width 4, attack 0.05 s, and release 0.2 s, respectively. As we use in iNeube-X 16 ms STFT windows, the length of the NAL-R equalization filter was set to 80 taps. When fed the oracle target anechoic speech, our STFT-domain NAL-R fitting plus compression algorithms produces an HASPI score of 0.987, which is slightly lower than the 0.998 score obtained by the CEC2 baseline fitting algorithm. This is likely due to the shorter NAL-R filter used in our algorithms, but our oracle score is still acceptable.

#### 1.5. System Training Configurations and External Data Usage

We trained models on two datasets, the original CEC2 data, which we scaled up to 16k scenes using the provided scene generation script, and this 16k scenes plus additional 6k scenes generated using the same configuration except for external clean speech taken from LibriVox in order to increase speaker variations.

Regarding training, we first train  $DNN_1$  with  $DNN_{\text{spk}}$  to convergence, and then freeze these two and train  $DNN_2$  using the output of these two plus the MCWF output. As explained in Section 1.1, the two MISO networks are trained to estimate the target anechoic signal at

Table 1: Results on development set of CEC2.

Approaches	SI-SDRi (dB)	STOI	PESQ	HASPI
Challenge baseline (mixture CH1 Left) [11]	0.0	0.563	1.197	0.245
Challenge baseline (oracle CH1 Left)	$\infty$	1.0	5.0	0.998
BLSTM-WPE+MVDR [12]	10.481	0.743	1.931	0.436
iNeuBe $DNN_1$ [2]	16.567	0.887	1.946	0.723
iNeuBe-X $DNN_1$	17.97	0.92	1.951	0.842
+ $DNN_2$ + NAL-R fine-tune	19.082	0.947	2.269	0.942
+ $DNN_2$ + NAL-R fine-tune + External Data	19.514	0.954	2.336	0.943

CH1 Left from the multi-channel mixture signal using complex spectral mapping. We use the Adam optimizer with 0.001 learning rate and batch size 1. We clip gradients with  $L_2$  norm more than 1. The MISO DNNs are trained on random 4 s segments of the input signal, which are constrained in such a way that the target and the interferer-only beginning part should appear both at least for 1 s. We use the whole development set for validation. The learning rate is halved if the validation loss is not improved over 5 epochs. We follow the scale-invariant loss function suggested in [2], which is defined on the estimated time-domain signal and its magnitude. Differently from [2], we use a multi-resolution STFT filterbanks with 512-, 1024-, 2048-, 256- and 128-sample windows to compute multiple magnitude losses. We observed that the magnitude loss leads to clearly better HASPI. Since this loss is scale-invariant, to prevent clipping in inference we re-scale the MISO DNNs’ estimates in a frame-online fashion using the output of the MCWF.

To further improve HASPI, after training  $DNN_2$  to convergence, we fine-tune  $DNN_2$  by adding an additional loss term where the same multi-resolution loss is computed between NAL-R fitted target anechoic speech and NAL-R fitted  $DNN$  estimates. Since listener audiograms are not available in the training set, we randomly sample audiograms from the development set listeners to perform this fitting-aware training.

## 2. RESULTS ON DEVELOPMENT SET

We report our results obtained on the CEC2 development set in Table 1 in terms of SI-SDR, STOI, PESQ and HASPI. Note that SI-SDR, STOI and PESQ are computed before listener adaptation and compression while HASPI is computed after. Together with the proposed iNeube-X, we also report the performance for two non-causal methods: iNeube (only  $DNN_1$ , based on TCNDenseNet) [2] and a DNN-based masked beamforming approach from the ESPNet-SE++ toolkit [12], which employs a two-layer bidirectional LSTM (BLSTM) and weighted prediction error (WPE) followed by minimum variance distortionless response (MVDR) beamforming. As another comparison, we report the scores of the unprocessed mixtures and the oracle target speech. The latter can be considered as a reasonable upper bound for our proposed system, since we employ a similar compensation strategy. Note that the CEC2 baseline system does not perform enhancement, and only performs listener adaptation and dynamic range compression. For our proposed systems, we report performance for various configurations and training strategies in the third panel: iNeube-X with only  $DNN_1$ , iNeube-X with  $DNN_2$  trained using the additional loss after listener adaptation (+  $DNN_2$  + NAL-R fine-tune), and with external data (+  $DNN_2$  + NAL-R fine-tune + External Data). The two latter systems are the ones we submitted to CEC2. We found that an additional iteration of beamforming followed by post-filtering led to slightly better HASPI, at a cost of increased computational overhead. Thus we only submitted signals obtained after the first  $DNN_2$  refinement step.

We emphasize that our system enjoys an algorithmic latency of 4 ms, and we confirm that the proposed system satisfies the 5 ms constraint on algorithmic latency using this public code<sup>1</sup>.

<sup>1</sup><https://github.com/zqwang7/CausalityCheck>

### 3. ACKNOWLEDGEMENTS

S. Cornell was partially supported by Marche Region within the funded project “Miracle” POR MARCHE FESR 2014-2020. Z.-Q. Wang used the Extreme Science and Engineering Discovery Environment [13], supported by the NSF under grant ACI-1548562, and the Bridges system [14], supported by the NSF under grant ACI-1445606, at the Pittsburgh Supercomputing Center. Y. Masuyama and N. Ono were partially supported by JSPS KAKENHI Grant Numbers JP21J21371 and JST CREST Grant Number JPMJCR19A3.

### 4. REFERENCES

- [1] Z.-Q. Wang, S. Cornell, S. Cho, Y. Lee, B.-Y. Kim, and S. Watanabe, “TF-GridNet: Making time-frequency domain models great again for monaural speaker separation,” in *submission to ICASSP 2023*, 2022.
- [2] Y.-J. Lu, S. Cornell, X. Chang, W. Zhang, C. Li, Z. Ni, Z.-Q. Wang, and S. Watanabe, “Towards low-distortion multi-channel speech enhancement: The ESPNET-SE submission to the L3DAS22 challenge,” in *Proc. ICASSP*, 2022, pp. 9201–9205.
- [3] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. L. Roux, “STFT-domain neural speech enhancement with very low algorithmic latency,” *arXiv preprint arXiv:2204.09911*, 2022.
- [4] Z.-Q. Wang, P. Wang, and D. Wang, “Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2001–2014, 2021.
- [5] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, “A consolidated perspective on multi-microphone speech enhancement and source separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 692–730, 2017.
- [6] Z.-Q. Wang, G. Wichern, and J. Le Roux, “Leveraging low-distortion target estimates for improved speech enhancement,” *arXiv preprint arXiv:2110.00570*, 2021.
- [7] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *Proc. ICASSP*, 2016, pp. 31–35.
- [8] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “FiLM: Visual reasoning with a general conditioning layer,” in *Proc. AAAI*, vol. 32, no. 1, 2018.
- [9] 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.
- [10] L. McCormack, V. Välimäki *et al.*, “FFT-based dynamic range compression,” in *Proc. Sound and Music Computing*, 2017, pp. 5–8.
- [11] X. Ren, L. Chen, X. Zheng *et al.*, “A neural beamforming network for B-Format 3D speech enhancement and recognition,” in *Proc. MLSP*, 2021.
- [12] Y.-J. Lu, X. Chang, C. Li, W. Zhang, S. Cornell, Z. Ni, Y. Masuyama, B. Yan, R. Scheibler, Z.-Q. Wang *et al.*, “ESPnet-SE++: Speech enhancement for robust speech recognition, translation, and understanding,” *Proc. Interspeech*, 2022.
- [13] J. Towns, T. Cockerill, M. Dahan *et al.*, “XSEDE: Accelerating scientific discovery,” *Computing in Science & Engineering*, vol. 16, no. 5, pp. 62–74, 2014.
- [14] N. A. Nystrom, M. J. Levine, R. Z. Roskies, and J. R. Scott, “Bridges: a uniquely flexible HPC resource for new communities and data analytics,” in *Proc. XSEDE*, 2015, pp. 1–8.