TOWARDS LOW-DISTORTION MULTI-CHANNEL SPEECH ENHANCEMENT: THE ESPNET-SE SUBMISSION TO THE L3DAS22 CHALLENGE

Yen-Ju Lu[1,5], *Samuele Cornell*[2], *Xuankai Chang*[1], *Wangyou Zhang*[4], *Chenda Li*[4], *Zhaoheng Ni*[3], *Zhong-Qiu Wang*[1], *Shinji Watanabe*[1]

[1]Carnegie Mellon University, USA [2]Università Politecnica delle Marche, Italy [3]Meta AI, USA
[4]Shanghai Jiao Tong University, Shanghai [5]Academia Sinica, Taipei

{cornellsamuele, neilyenjulu, nizhaoheng, wang.zhongqiu41}@gmail.com

ABSTRACT

This paper describes our submission to the L3DAS22 Challenge Task 1, which consists of speech enhancement with 3D Ambisonic microphones. The core of our approach combines Deep Neural Network (DNN) driven complex spectral mapping with linear beamformers such as the multi-frame multi-channel Wiener filter. Our proposed system has two DNNs and a linear beamformer in between. Both DNNs are trained to perform complex spectral mapping, using a combination of waveform and magnitude spectrum losses. The estimated signal from the first DNN is used to drive a linear beamformer, and the beamforming result, together with this enhanced signal, are used as extra inputs for the second DNN which refines the estimation. Then, from this new estimated signal, the linear beamformer and second DNN are run iteratively. The proposed method was ranked first in the challenge, achieving, on the evaluation set, a ranking metric of 0.984, versus 0.833 of the challenge baseline.

Index Terms— beamforming, multi-microphone complex spectral mapping, multi-channel speech enhancement, deep learning.

1. INTRODUCTION

Multi-channel speech enhancement (SE) is an important pre-processing step for many applications, such as hands-free speech communication, hearing aids, smart speakers, and automatic speech recognition (ASR) [1]. In its broad definition, it consists of joint denoising and dereverberation of a desired target speech signal from a noisy-reverberant multi-channel mixture signal captured by a microphone array. This arduous problem has been effectively addressed in the last decade with DNN-based methods, which have been firmly established as the de-facto mainstream approach for speech enhancement [2]. Multi-channel DNN-based methods can be roughly divided into *hybrid* [3–7] and *fully-neural* [8–12] methods. The former combines DNNs with conventional signal processing based techniques, using the DNNs to drive, for example, dereverberation algorithms such as Weighted Prediction Error (WPE) [13] or classical beamforming algorithms such as Minimum Variance Distortionless Response (MVDR) and Multi-Channel Wiener Filter (MCWF) [1, 14]. In fully-neural systems, DNNs are trained to directly estimate the target speech from the mixture. The DNN can have either a Multiple Input Single Output (MISO) structure or can be used to directly estimate linear beamforming filters in the Short-Time Fourier Transform (STFT) domain [9] or in the time domain [8]. Fully-neural methods are effective and often outperform hybrid techniques for what regards signal-level SE metrics such as Scale Invariant Signal-to-Distortion Ratio (SI-SDR) [15],

Short-time Objective Intelligibility (STOI) [16] etc. However, unlike conventional dereverberation and beamforming algorithms, they tend to introduce non-linear distortions, which can degrade the performance of downstream tasks, such as Automatic Speech Recognition (ASR) [17, 18]. This problem can be mitigated by using end-to-end training or Deep Feature Loss (DFL) [19, 20]. On the other hand, these techniques requires re-training or fine-tuning whenever the back-end model changes and are cumbersome to apply when there are multiple downstream tasks.

This trade-off between signal-level SE metrics and ASR performance is at the core of the L3DAS22 speech enhancement challenge since the models are ranked by considering both STOI [21] and Word Error Rate (WER). While correlated to some degree, the two metrics reflect two highly different downstream application scenarios: ASR and human listening for STOI. Our goal is to devise a "generalist" SE model, optimized independently from these metrics or the backend ASR models, but able to significantly improve both.

To address this arduous problem, we employ a framework derived from previous works [22, 23], which combines the merits of hybrid and fully-neural methods: namely, beamforming's ability at producing low-distortion estimates and DNN's high capacity at suppressing non-target signals. Compared with [22], in this study we perform multi-channel enhancement by beamforming directly on 3D Ambisonic microphone format. We introduce here one main novelty: applying a multi-frame beamformer [7, 24, 25] with the beamforming filter estimated directly from the DNN estimated target signal. We show that this helps to tackle the problem of target signal misalignment explained in Section 2. We propose an iterative neural/beamforming enhancement (iNeuBe) architecture inculding two TCN-DenseUNet [23] which are employed in a MISO configuration and a beamformer. Our system is depicted in Fig. 1. The first DNN (DNN_1) takes in input the complex STFT coefficients of the multi-channel input mixture signal ($\mathbf{Y}$) and regresses directly the complex STFT coefficients of the target signal ($\hat{S}^{(1)}$). DNN_1 enhanced signal ($\hat{S}^{(1)}$) is used to drive a multi-frame MCWF (mfMCWF) at the first ($i = 0$) iteration to derive a low-distortion estimate of the target signal ($\hat{S}_{i=0}^{(\text{mfMCWF})}$). Both $\hat{S}_{i=0}^{(\text{mfMCWF})}$ and $\hat{S}^{(1)}$ are then used as additional features for the second DNN (DNN_2) to refine the target estimate. The output of DNN_2 ($\hat{S}_{i=0}^{(2)}$) can be used iteratively in place of $\hat{S}^{(1)}$ to compute another refined beamforming result $\hat{S}_{i=1}^{(\text{mfMCWF})}$ which is then fed back to DNN_2 together with $\hat{S}_{i=0}^{(2)}$.

The proposed framework placed first in the L3DAS22 speech enhancement challenge, achieving a Task 1 challenge metric of 0.984 on the evaluation set, versus 0.833 achieved by the official baseline and 0.975 by the runner-up system. This indicates that the proposed

approach is a promising step towards "generalist" multi-channel SE, as it achieves both low WER and high STOI without any fine-tuning with the back-end ASR model or use of STOI-derived losses.

We have made our implementation available through the ESPNet-SE toolkit [20].

2. L3DAS22 TASK 1 DESCRIPTION

The L3DAS22 3D speech enhancement task (Task 1) [21] challenges participants to predict the dry speech source signal from its far-field mixture recorded by two four-channel Ambisonic-format signals in a noisy-reverberant office environment. The challenge dataset is "semi-synthetic". It consists of 252 measured room impulse responses (RIRs). The dry speech source signals are sampled from LibriSpeech [26] and the dry noise signals from FSD50K [27]. Two first-order A-format Ambisonic microphones arrays, each with four microphones, are employed to record the RIRs. A single room is used for RIR measurement. The microphone placement is fixed, with one at the room center and the other 20 cm apart. Notably, the room configuration and microphone placement do not change between training and testing, and the source positions are distributed uniformly inside the room. Artificial mixtures are generated by convolving dry speech and noise signals with the measured RIRs and mixing the convolved signals together. The Signal-to-Noise-Ratio (SNR) is sampled from the range [6, 16] dBFS (decibels relative to full scale). The generated A-format Ambisonic mixtures are then converted to B-format Ambisonic. The total amount of data is around 80 hours for training and 6 hours for development.

The challenge ranks the submitted systems using a combination of STOI [28] and WER:

$$\text{Task1 Metric} = \big(\text{STOI} + (1 - \text{WER}^{\perp 1})\big)/2. \qquad (1)$$

where $\text{WER}^{\perp 1} = \min(\text{WER}, 1)$. The values of STOI and WER are both in the range of $[0, 1]$, so is the composite metric. The WER is computed based on the transcription of the estimated target signal and that of the reference signal, both decoded by a pre-trained Wav2Vec2 ASR model [29].

We emphasize that the goal of the challenge is recovering the dry speech signal from a far-field noisy-reverberant mixture. As such, the metrics above are computed with respect to the dry ground truth. This makes the task extremely challenging, because, besides removing reverberation and noises, a system also needs to time-align the estimated signal with the dry speech signal in order to obtain a good STOI. STOI, in fact, contrary to WER, is highly sensitive to time-shifts: e.g. a shift in the order of 100 samples alone can decrease the STOI value from 1.0 to 0.9 for the very same oracle target speech. Thus it is required that the model performs, either implicitly or explicitly, localization of the target source inside the room, so that an aligned estimate can be produced.

3. PROPOSED METHOD

3.1. System Overview

Let us denote the dry speech source signal as $s[n] \in \mathbb{R}$ and the far-field mixture recorded by Ambisonic microphones as $\mathbf{y}[n] \in \mathbb{R}^P$, where n indexes discrete time and P ($= 8$ in this study) is the number of channels. Following the challenge baseline [9], our proposed system operates on the STFT spectra of the B-format Ambisonic signals. We denote the STFT coefficients of the mixture and dry speech signal at time t and frequency f as $\mathbf{Y}(t, f) \in \mathbb{C}^P$ and $S(t, f) \in \mathbb{C}$, respectively. For simplicity, we will omit in the following the t and
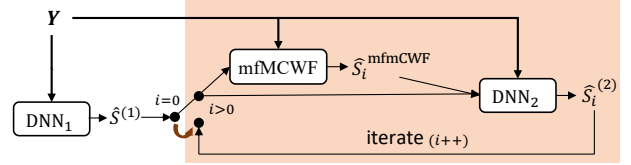
Fig. 1: Overview of proposed iterative neural/beamforming enhancement (iNeuBe) framework. A multi-frame multi-channel Wiener filter (mfMCWF) beamformer is applied between the two DNN MISO networks.

f indexes, and denote the STFT spectra simply as $\mathbf{Y}$ and S, and signals as $\mathbf{y}$ and s.

Our proposed iNeuBe framework, illustrated in Fig. 1, contains two DNNs and a linear beamforming module in between. Both DNNs have a MISO structure and are trained using multi-microphone complex spectral mapping [12, 30, 31], where the real and imaginary (RI) components of multiple input signals are concatenated as input features for the DNNs to predict the RI components of the target speech.

More in detail, for DNN_1 we concatenate the RI components of $\mathbf{Y}$ as input to predict the RI components of S. DNN_1 produces an estimated target speech $\hat{S}^{(1)}$, which is at the first iteration $i = 0$ used to compute an mfMCWF for the target speech. Subsequently, the RI components of the beamforming result $\hat{S}_{i=0}^{\text{mfMCWF}}$, the input mixture $\mathbf{Y}$, and $\hat{S}^{(1)}$ are concatenated and fed as input for DNN_2 to further refine the estimation for the RI components of S. DNN_2 produces another refined estimation of S, i.e. $\hat{S}_{i=0}^{(2)}$, which can be used iteratively in place of $\hat{S}^{(1)}$ to recompute the beamforming result and as an additional feature to DNN_2.

The rest of this section describes the DNN architecture, the loss function employed for the DNN training, the mfMCWF beamforming algorithm, and the run-time iterative procedure.

3.2. Multi-Microphone Complex Spectral Mapping

We employ the TCN-DenseUNet architecture described in the Fig. 15 of [23] for both DNN_1 and DNN_2 (the parameters are not shared). It is a temporal convolution network (TCN) sandwiched by a U-Net derived structure. DenseNet blocks are inserted at multiple frequency scales of the encoder and decoder of the U-Net. This network architecture has shown strong performance in tasks such as speech enhancement, dereverberation and speaker separation [12, 22, 23]. The network takes as input a real-valued tensor with shape $C \times T \times F$, where C is the number of channels, T the number of STFT frames and F the number of STFT frequencies. The RI components of different input signals are concatenated along the channel axis and fed as feature maps to the network. In the case of DNN_1, C equals 16 as we have 8 microphone channels. Linear activation units are used in the output layer to obtain the predicted RI components of the target signal. Each network has around 6.9 million parameters.

Given the DNN-estimated RI components, denoted as $\hat{R}^{(b)}$ and $\hat{I}^{(b)}$ where $b \in \{1, 2\}$ indicates which of the two DNNs produces the outputs, we compute the enhanced speech as $\hat{S}^{(b)} = \hat{R}^{(b)} + j\hat{I}^{(b)}$, where j is the imaginary unit, and use inverse STFT (iSTFT) to re-synthesize the time-domain signal $\hat{s}^{(b)} = \text{iSTFT}(\hat{S}^{(b)})$. After that, we equalize the gains of the estimated and true source signals by using a scaling factor $\ddot{\alpha}$, and define the loss function on the scaled, re-synthesized signal and its STFT magnitude:

$$\mathcal{L}_{\text{Wav+Mag}}^{(b)} = \|\ddot{\alpha}\hat{s}^{(b)} - s\|_1 + \Big\| |\text{STFT}(\ddot{\alpha}\hat{s}^{(b)})| - |\text{STFT}(s)| \Big\|_1, \quad (2)$$

where $\|\cdot\|_1$ calculates the L_1 norm, $|\cdot|$ computes magnitude, and STFT$(\cdot)$ extracts a complex spectrogram. $\ddot{\alpha} = \text{argmin}_\alpha \|\alpha \hat{s}^{(b)} - s\|_2^2 = (s^\mathsf{T} \hat{s}^{(b)})/(\hat{s}^{(b)\mathsf{T}} \hat{s}^{(b)})$, where $\|\cdot\|_2$ computes the L_2 norm. The loss on magnitude can improve metrics such as STOI and WER which favor signals with a good estimated magnitude [32].

Before training, we normalize the sample variance of the multi-channel input mixture to 1.0. We do the same normalization also for the dry speech source signal. We found this normalization procedure essential for training our models as there is a significant gain mismatch between the reference and the mixture signals.

3.3. Multi-Frame MCWF

Based on the estimated target signal $\hat{S}^{(b)}$ produced by DNN$_1$ or DNN$_2$, we compute an mfMCWF per frequency through the following minimization problem:

$$\min_{\mathbf{w}(f)} \sum_t \left| \hat{S}^{(b)}(t,f) - \mathbf{w}(f)^\mathsf{H} \widetilde{\mathbf{Y}}(t,f) \right|^2, \quad (3)$$

where $\widetilde{\mathbf{Y}}(t,f) = [\mathbf{Y}(t-l,f)^\mathsf{T}, \ldots, \mathbf{Y}(t,f)^\mathsf{T}, \ldots, \mathbf{Y}(t+r,f)^\mathsf{T}]^\mathsf{T}$ and $\mathbf{w}(f) \in \mathbb{C}^{(l+1+r)P}$. l and r control the context of frames for beamforming, leading to a single-frame MCWF when l and r are zeros, and an mfMCWF when l and r are positive. The minimization problem is quadratic, and a closed-form solution $\hat{\mathbf{w}}(f)$ is available:

$$\hat{\mathbf{w}}(f) = \hat{\boldsymbol{\Phi}}(f)^{-1} \hat{\mathbf{z}}^{(b)}(f) \quad (4)$$

$$\hat{\boldsymbol{\Phi}}(f) = \sum_t \widetilde{\mathbf{Y}}(t,f) \widetilde{\mathbf{Y}}(t,f)^\mathsf{H} \quad (5)$$

$$\hat{\mathbf{z}}^{(b)}(f) = \sum_t \widetilde{\mathbf{Y}}(t,f) \hat{S}^{(b)}(t,f)^*, \quad (6)$$

where $(\cdot)^*$ computes complex conjugate. The beamforming result $\hat{S}^{\text{mfMCWF}}$ is computed as:

$$\hat{S}^{\text{mfMCWF}} = \hat{\mathbf{w}}(f)^\mathsf{H} \widetilde{\mathbf{Y}}(t,f). \quad (7)$$

Notice that in the computation of $\hat{\mathbf{z}}^{(b)}(f)$ and $\hat{\boldsymbol{\Phi}}(f)$, we average over all the frames in each utterance and compute a time-invariant beamformer, implicitly assuming that the transfer functions between the arrays and sources do not change within each utterance. This is a valid assumption for the L3DAS22 setup [21]. We emphasize that our approach directly performs beamforming on Ambisonic signals.

As outlined in Section 2, the dry source signal is not time-aligned with the far-field mixture. In this scenario, a multi-frame beamformer is highly desirable, as a larger context of frames can be leveraged by the beamformer to compensate the signal shift. This DNN-supported mfMCWF was proposed recently in [7]. The major difference is that here we use multi-microphone complex spectral mapping to obtain $\hat{S}^{(b)}$, which consists of DNN-estimated magnitude and phase. In contrast, [7] uses monaural real-valued magnitude masking on the far-field mixture to obtain $\hat{S}^{(b)}$ and hence $\tilde{S}^{(b)}$ has the mixture phase. When target speech is not time-aligned with the mixture, our approach is clearly more principled, as the DNN is free to estimate an $\hat{S}^{(b)}$ that is sample-aligned with S. Instead, if real-valued masking is used, the estimated signal would be aligned with the mixture.

For similar reasons, other multi-frame filters [24, 25] cannot align their predictions with the dry target signal. In addition, although they have shown good performance for signals recorded by omnidirectional microphones, it is not clear whether they can be directly applied for signals in Ambisonic format. In contrast, our mfMCWF can readily deal with both formats, without any changes.

Table 1: Results of one-DNN systems on dev. set. Approaches marked with * use additional STOI loss and ASR-based Deep Feature loss.

Approaches	WER (%)	STOI	Task1 Metric
Challenge Baseline [9]	25.0	0.870	0.810
FasNet* [8]	18.2	0.874	0.846
Conv-TasNet [36] MVDR*	5.56	0.821	0.883
DCCRN* [33]	18.8	0.907	0.860
Demucs v2* [34]	26.3	0.851	0.794
Demucs v3* [38]	15.3	0.874	0.860
DNN$_1$	**3.90**	**0.964**	**0.963**

3.4. Run-Time Iterative Processing

Ar run time, we can iterate the orange block in Fig. 1 to gradually refine the target estimate. Denoting as $\hat{S}^{(2)}_{i=0}$ the output of DNN$_2$ after the first pass, we can use this estimate in place of $\hat{S}^{(1)}$ to run again the beamforming module and obtain $\hat{S}^{\text{mfMCWF}}_{i=1}$. This new beamformed estimate, together with $\hat{S}^{(2)}_{i=0}$ and $\mathbf{Y}$, can then be fed back again to DNN$_2$ to produce an estimate at iteration two, $\hat{S}^{(2)}_{i=1}$, and so on.

4. EXPERIMENTAL SETUP

4.1. Configurations of Proposed Method

Regarding our iNeuBe architecture, we use an STFT window size of 32 ms and a hop size of 8 ms. As analysis window we employ square-root Hanning. DNN$_1$ and DNN$_2$ are trained separately and in a sequential manner: after DNN$_1$ is trained, we run it on the entire training set to generate the beamforming results and $\hat{S}^{(1)}$, and then train DNN$_2$ using the generated signals as the extra input. Regarding the mfMCWF module, we compare the performance of setting l and r to different values in Section 5. Regarding the challenge results, we use the immediate outputs from the model without any post-processing.

4.2. Benchmark Systems

In this challenge, we also experimented with several state-of-the-art enhancement models such as DCCRN [33], Demucs v2 and v3 [34], and FasNet [8]. We also explored an improved version of the mask-based beamforming model used in [35], which is based on ConvTas-Net [36]. This model is directly derived from [35] and uses ConvTas-Net separator to predict a magnitude mask for the target signal. The mask is then used to derive a time-invariant MVDR filter which is employed to estimate the target speech. Differently from [35], here we employ TAC [37] after every repeat in the ConvTasNet separator (we use the standard values of $R = 3$ repeats and $X = 8$ blocks). DCCRN, Demucs v2 and v3 systems instead rely on complex spectral mapping, without explicit beamforming operations.

5. RESULTS

Table 1 compares the challenge metrics obtained by the different models introduced in Section 4.2. For these models we used additional losses related to the challenge metrics: namely the STOI loss and a DFL derived from the Wav2Vec2 ASR back-end used by the challenge to compute the WER scores. In detail we used as DFL the log-Mean-Squared Error (MSE) between the Wav2Vec2 final-layer activations when it is fed the enhanced signal versus when it is fed the oracle target speech signal. Despite the proposed model is trained in a back-end agnostic way, i.e. without using DFL and STOI

Table 2: Comparison of various beamforming setup on dev. set.

Approaches	l	r	WER (%)	STOI	Task1 Metric
DNN_1	-	-	3.90	0.964	0.963
DNN_1+mfMCWF	0	0	6.98	0.917	0.923
DNN_1+mfMCWF	7	0	3.42	0.966	0.966
DNN_1+mfMCWF	6	1	3.13	0.974	0.971
DNN_1+mfMCWF	5	2	3.09	0.974	0.972
DNN_1+mfMCWF	4	3	**3.04**	**0.975**	**0.972**
Magnitude-mask based mfMCWF [7]	4	3	4.82	0.959	0.955

Table 3: Results of two-DNN systems on dev. set.

Approaches	l	r	WER (%)	STOI	Task1 Metric
Challenge Baseline [9]	-	-	25.0	0.870	0.810
DNN_1	-	-	3.90	0.964	0.963
DNN_1+MVDR+DNN_2	-	-	3.62	0.970	0.968
DNN_1+mfMCWF+DNN_2	0	0	3.36	0.971	0.969
DNN_1+mfMCWF+DNN_2	7	0	2.63	0.978	0.976
DNN_1+mfMCWF+DNN_2	6	1	2.36	0.982	0.979
DNN_1+mfMCWF+DNN_2	5	2	2.53	0.982	0.978
DNN_1+mfMCWF+DNN_2	4	3	2.35	0.983	0.980
DNN_1+(mfMCWF+DNN_2)×2	4	3	**2.14**	**0.986**	**0.982**

Table 4: Results of two-DNN systems on eval. set.

Approaches	l	r	WER (%)	STOI	Task1 Metric
DNN_1	-	-	3.73	0.964	0.964
DNN_1+mfMCWF+DNN_2	0	0	3.15	0.971	0.970
DNN_1+mfMCWF+DNN_2	7	0	2.28	0.978	0.978
DNN_1+mfMCWF+DNN_2	4	3	2.11	0.983	0.981
DNN_1+(mfMCWF+DNN_2)×2	4	3	**1.89**	**0.987**	**0.984**
Challenge baseline [9]	-	-	21.2	0.878	0.833
Runner-up system (BaiduSpeech)	-	-	2.50	0.975	0.975

Table 5: Results of using one Ambisonic microphone on dev. set.

Approaches	l	r	WER (%)	STOI	Task1 Metric
DNN_1	-	-	4.11	0.958	0.958
DNN_1+mfMCWF+DNN_2	12	3	2.45	0.980	0.978
DNN_1+(mfMCWF+DNN_2)×2	12	3	**2.49**	**0.982**	**0.979**

related losses, it significantly outperforms the other models which instead rely on additional loss terms associated with the particular challenge task. In addition, a noticeable trend is that the models employing complex spectral mapping (DNN_1, DCCRN and Demucs v2 and v3) consistently obtain higher STOI than Conv-TasNet MVDR, which is based on mask-based beamforming. The models that rely on complex spectral mapping, being unconstrained, are capable of producing an aligned estimate with respect to the true oracle signal, leading to inherently higher STOI. In contrast, mask-based beamforming methods, as explained in Section 3, produce an estimate that is constrained to be aligned with the input mixture signals.

In Table 2, we first report the mfMCWF results by using DNN_1's output to compute the beamformer. We set l and r to different values. We can see that mfMCWF consistently outperforms single-frame MCWF, which is the same as mfMCWF with $l = 0$ and $r = 0$. The best performance is obtained by using a quasi-symmetrical configuration of $l = 4$ past frames and $r = 3$ future frames, and the resulting linear mfMCWF even obtains better scores than the non-linear DNN_1. For comparison, we also report the result of the magnitude-mask based mfMCWF in [7]. In this latter model, we slightly modify the TCN-DenseUNet architecture, and train through the mask based mfMCWF and compute the loss in Eq. (2) based on the beamformed signal. We found this training-through mechanism essential to make the mask-based mfMCWF work. We tried using the DNN_1's output to derive a magnitude mask and compute the beamformer (i.e. without training-through). However, this led to severely degraded performance, because the DNN_1's output is not time- and gain-aligned with the far-field mixture and hence it is not straightforward how to compute a valid magnitude mask. Also, the mask based mfMCWF needs to designate one of the microphones as the reference, meaning that the resulting beamformed signal cannot be fully aligned with the dry source signal. For this reason, complex spectral mapping for mfMCWF computation leads to higher performance.

In Table 3 we report the results of including DNN_2 into our system. Clear improvement is obtained over DNN_1 and DNN_1+mfMCWF. Run-time iterative estimation (up to two iterations), denoted as DNN_1+(mfMCWF+DNN_2)×2, further improves the performance, at a cost of increased computational requirements.

In Table 4 we report the results obtained on the challenge evaluation set by a subset of configurations explored in Table 3. We notice that the results between the development set and evaluation set are consistent. Our proposed approach ranked first among all the sub-

missions[1] to the L3DAS22 Task 1 speech enhancement challenge and shows a remarkable improvement over the baseline system and a significant improvement over the runner-up system.

In Table 5 we additionally provide the results obtained by only using the first ambisonic microphone for testing. The signals at both ambisonic microphones are used for training, and this doubles the number of training examples. The number of filter taps for mfMCWF is increased from 8 to 16. The results on the development set are close to the ones obtained by using both ambisonic microphones for training and testing (compare the last rows of Table 5 and 3).

6. CONCLUSIONS

In this paper we have described our submission to the L3DAS22 Task 1 challenge. Our proposed iNeuBe framework relies on an iterative pipeline of linear beamforming and DNN-based complex spectral mapping. In our method, two DNNs are employed in a MISO configuration and use complex spectral mapping to estimate the target speech signal. The first DNN output is used to drive an mfMCWF, and a second DNN, taking the outputs of the first DNN and the mfMCWF as additional input features, is used to further refine the estimated target speech signal from the first DNN. The second DNN and linear beamforming can be run iteratively and we show that up to the second iterations there are noticeable improvements, especially regarding WER.

Compared to previous work, we propose here the use of mfMCWF and show that computing mfMCWF weights using DNN-based complex spectral mapping output can have significant advantages in the challenge scenario. Our proposed method ranked first in the L3DAS22 challenge, significantly outperforming the baseline and the second-best system. As additional contributions we also performed several ablation studies weighting different configurations and the contribution of each block in the iNeuBe framework.

Finally, we also compared our proposed approach with multiple state-of-the-art models and showed that it can achieve remarkably better challenge metrics, with both lower WER and higher STOI, even when the competing models are trained with back-end task aware losses.

7. ACKNOWLEDGEMENTS

---

[1]See https://www.l3das.com/icassp2022/results.html for the full ranking.

8. REFERENCES

[1] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[3] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. ASRU*, 2015.

[4] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *Proc. ICASSP*, 2018.

[5] H. Erdogan, J. R. Hershey, S. Watanabe *et al.*, "Improved MVDR beamforming using single-channel mask prediction networks." in *Proc. Interspeech*, 2016.

[6] T. Ochiai, S. Watanabe, T. Hori *et al.*, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017.

[7] Z.-Q. Wang, H. Erdogan, S. Wisdom *et al.*, "Sequential multi-frame neural beamforming for speech separation and enhancement," in *Proc. SLT*, 2021.

[8] Y. Luo, C. Han, N. Mesgarani *et al.*, "FasNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *Proc. ASRU*, 2019.

[9] X. Ren, L. Chen, X. Zheng *et al.*, "A neural beamforming network for B-Format 3D speech enhancement and recognition," in *Proc. MLSP*, 2021.

[10] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense U-Net for multichannel speech enhancement," in *Proc. ICASSP*, 2020.

[11] C.-L. Liu, S.-W. Fu, Y.-J. Li *et al.*, "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1888–1900, 2020.

[12] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2001–2014, 2021.

[13] K. Kinoshita, M. Delcroix, H. Kwon *et al.*, "Neural network-based spectrum estimation for online WPE dereverberation," in *Proc. Interspeech*, 2017.

[14] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016.

[15] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *Proc. ICASSP*, 2019.

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

[17] K. Iwamoto, T. Ochiai, M. Delcroix *et al.*, "How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR," *arXiv preprint arXiv:2201.06685*, 2022.

[18] W. Zhang, J. Shi *et al.*, "Closing the gap between time-domain multi-channel speech enhancement on real and simulation conditions," in *Proc. WASPAA*, 2021, pp. 146–150.

[19] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, "Spectral feature mapping with mimic loss for robust speech recognition," in *Proc. ICASSP*, 2018, pp. 5609–5613.

[20] C. Li, J. Shi, W. Zhang *et al.*, "ESPnet-SE: end-to-end speech enhancement and separation toolkit designed for ASR integration," in *Proc. SLT*, 2021.

[21] E. Guizzo, C. Marinoni, M. Pennese *et al.*, "L3DAS22 challenge: Learning 3D audio sources in a real office environment," in *Proc. ICASSP*, 2022.

[22] Z.-Q. Wang, G. Wichern, and J. Le Roux, "Convolutive prediction for monaural speech dereverberation and noisy-reverberant speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3476–3490, 2021.

[23] ——, "Leveraging low-distortion target estimates for improved speech enhancement," *arXiv preprint arXiv:2110.00570*, 2021.

[24] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Processing Letters*, vol. 26, no. 6, pp. 903–907, 2019.

[25] Z. Zhang, Y. Xu *et al.*, "Multi-Channel Multi-Frame ADL-MVDR for Target Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3526–3540, 2021.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015.

[27] E. Fonseca, X. Favory, J. Pons *et al.*, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2021.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010.

[29] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[30] Z.-Q. Wang and D. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," in *Proc. ICASSP*, 2020, pp. 486–490.

[31] K. Tan, Z.-Q. Wang, and D. Wang, "Neural spectrospatial filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 605–621, 2022.

[32] Z.-Q. Wang, G. Wichern, and J. Le Roux, "On the compensation between magnitude and phase in speech separation," *IEEE Signal Process. Lett.*, vol. 28, pp. 2018–2022, 2021.

[33] Y. Hu, Y. Liu, S. Lv *et al.*, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," *Proc. Interspeech*, 2020.

[34] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Demucs: Deep extractor for music sources with extra unlabeled data remixed," *arXiv preprint arXiv:1909.01174*, 2019.

[35] S. Cornell, M. Pariente, F. Grondin, and S. Squartini, "Learning filterbanks for end-to-end acoustic beamforming," *arXiv e-prints*, pp. arXiv–2111, 2021.

[36] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, p. 1256–1266, 2019.

[37] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. ICASSP*, 2020.

[38] A. Défossez, "Hybrid spectrogram and waveform source separation," *Proc. ISMIR*, 2021.

[39] J. Towns, T. Cockerill, M. Dahan *et al.*, "XSEDE: Accelerating scientific discovery," *Computing in Science & Engineering*, vol. 16, no. 5, pp. 62–74, 2014.

[40] 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.