iNeuBe: Towards Low-distortion Multi-channel Speech Enhancement

Yen-Ju Lu, Samuele Cornell, Xuankai Chang, Wangyou Zhang, Chenda Li,
Zhaoheng Ni, Zhong-Qiu Wang, Shinji Watanabe

Agenda

- Introduction
- iNeuBe Framework
- Empirical Results on L3DAS22
- Conclusions

Multi-channel Speech Enhancement

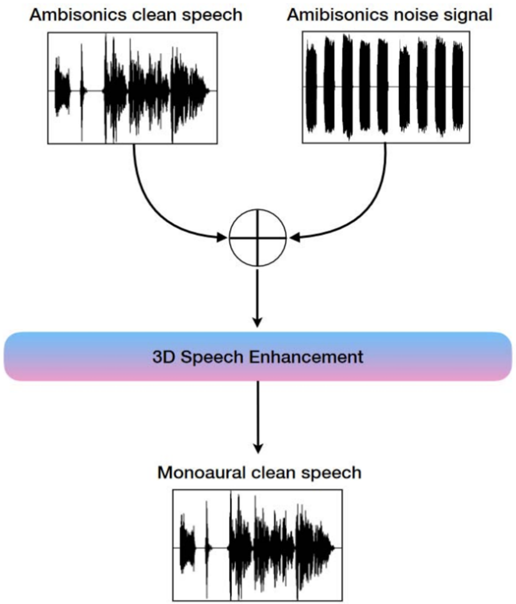
- Aims at estimating clean speech from audio recordings by multiple microphones.
- Given multi-channel noisy reverberant mixture speech, the Short-Time Fourier Transform (STFT) coefficients of mixture Y can be modeled as:

$$\mathbf{Y}(t, f) = \mathbf{S}(t, f) + \mathbf{H}(t, f) + \mathbf{N}(t, f)$$

  - where S(t, f), H(t, f), N(t, f) denote the STFT vectors of the direct and non-direct signals of the target speaker, reverberant noise, respectively the at time t and frequency f.
  - S(t, f) + H (t, f) is the reverberant speech of target speaker.
- Task difficulty varies depending on the target.

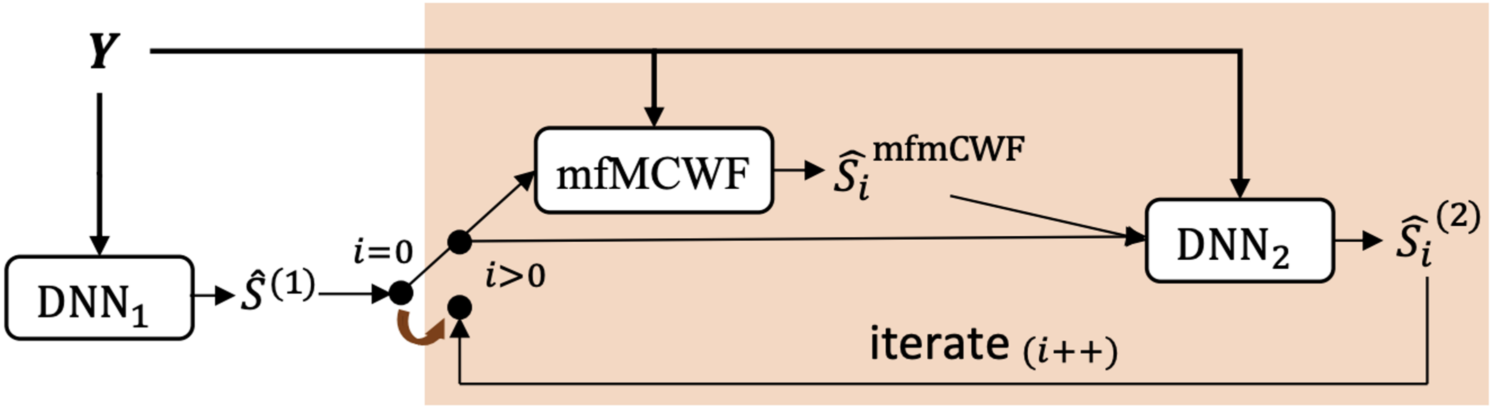
L3DAS22 Challenge

- Multi-channel mixture data
  - 8-channels 16kHz wav files.
  - 2 sets of first-order* B-Format Ambisonics microphone array
  - *first-order == 4 channels
- Target
  - Single-channel dry-clean speech (w/o reverberation).
- Evaluation metric:
  - Score = (STOI + (1 - WER)) / 2
  - WER* is computed by a pre-trained Wav2Vec2 ASR model.
    - *WER(hypo_clean, hypo_estimate)



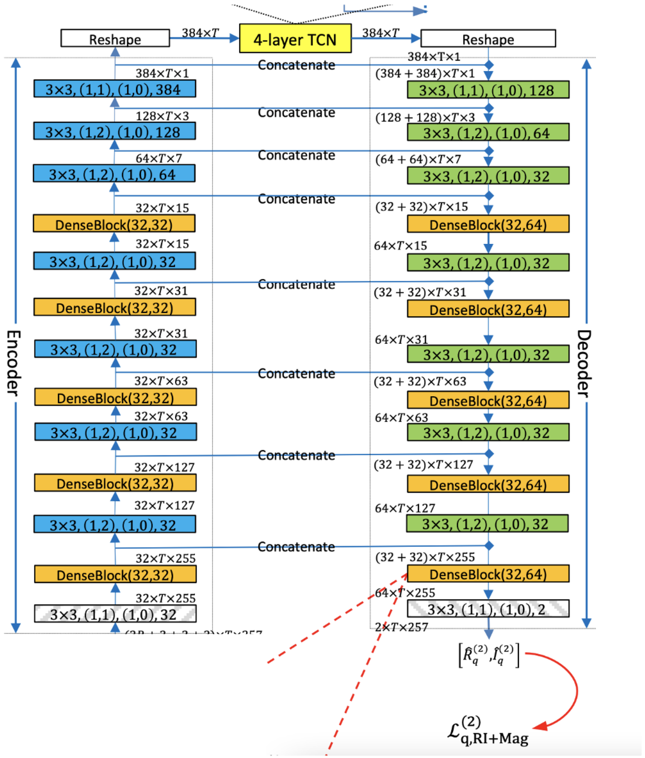
L3DAS challenge: https://www.l3das.com/icassp2022/task1.html

iNeuBe: iterative Neural Beamforming Enhancement

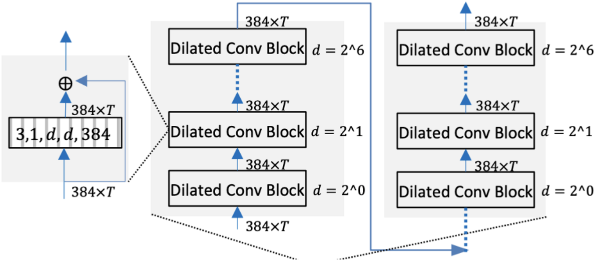


- Estimate enhanced Real + Imaginary components $S^{(1)}$ via DNN$_1$
- Use $S^{(1)}$ as target for Multi-frame Multi-channel Wiener Filter (mfMCWF)
- Use $S^{(1)}$ and $S^{(mfMCWF)}$ as input feature to estimate $S^{(2)}$ via DNN$_2$

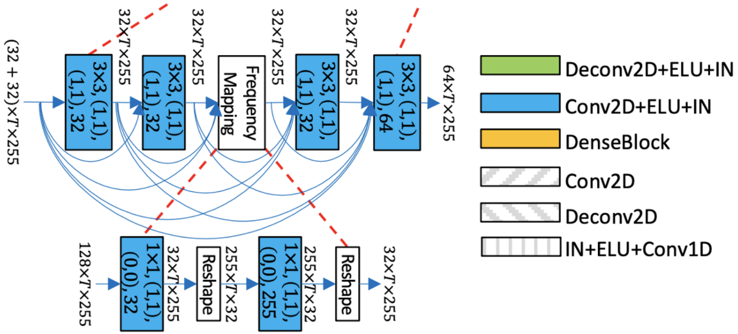
DNN Architecture: TCN-DenseUNet



(temporal convolutional network) TCN



Dense Block



Wang, Zhong-Qiu, Gordon Wichern, and Jonathan Le Roux. "Leveraging Low-Distortion Target Estimates for Improved Speech Enhancement." *arXiv preprint arXiv:2110.00570* (2021).

Multi-frame MCWF

- Based on the estimated target signal $S^{(b)}$ produced by DNN_1 or DNN_2, we compute the mfMCWF weight per frequency by optimizing

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

  - Where $\widetilde{\mathbf{Y}}(t,f) = [\mathbf{Y}(t-l,f)^{\mathsf{T}}, \dots, \mathbf{Y}(t,f)^{\mathsf{T}}, \dots, \mathbf{Y}(t+r,f)^{\mathsf{T}}]^{\mathsf{T}}$ and $\mathbf{w}(f) \in \mathbb{C}^{(l+1+r)P}$
  - l and r controls the history frame and future frame indices, respectively.
  - P denotes the number of channels.
  - Set l and r to 0 leads to **single-frame** MCWF.
- The beamforming output is computed as:

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

Baseline Systems

- Official baseline[1]
  - UNet + beamforming
- FasNet[2]
- Multi-channel Conv-TasNet[3] + MVDR beamforming
- DCCRN[4]
- Demucs v2[5]
- Demucs v3[6]

1 Ren, Xinlei, Lianwu Chen, Xiguang Zheng, Chenglin Xu, Xu Zhang, Chen Zhang, Liang Guo, and Bing Yu. "A Neural Beamforming Network for B-Format 3D Speech Enhancement and Recognition." In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1-6. IEEE, 2021.

2 Luo, Yi, Cong Han, Nima Mesgarani, Enea Ceolini, and Shih-Chii Liu. "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing." In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pp. 260-267. IEEE, 2019.

3 Luo, Yi, and Nima Mesgarani. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation." *IEEE/ACM transactions on audio, speech, and language processing* 27, no. 8 (2019): 1256-1266.

4 Hu, Yanxin, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement." *arXiv preprint arXiv:2008.00264* (2020).

5 Défossez, Alexandre, Nicolas Usunier, Léon Bottou, and Francis Bach. "Demucs: Deep extractor for music sources with extra unlabeled data remixed." *arXiv preprint arXiv:1909.01174* (2019).

6 Défossez, Alexandre. "Hybrid Spectrogram and Waveform Source Separation." *arXiv preprint arXiv:2111.03600* (2021).

Loss Function

- After computing the RI components $S^{(b)}$ ($S^{(1)}$, $S^{(2)}$, or $S^{(mfMCWF)}$), compute the waveforms by an iSTFT layer.

$$\hat{s}^{(b)} = \mathsf{iSTFT}(\hat{S}^{(b)})$$

- The loss is the combination of L1 losses on waveforms and magnitudes, respectively.

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

  - where $\ddot{\alpha} = \mathsf{argmin}_\alpha \|\alpha\hat{s}^{(b)} - s\|_2^2 = \left(s^\mathsf{T}\hat{s}^{(b)}\right)/\left(\hat{s}^{(b)^\mathsf{T}}\hat{s}^{(b)}\right)$ is the scaling factor

Additional Losses for Baselines

- STOI loss
  - Compute the log(STOI)
  - Back-propagate the gradient
- ASR-based Deep Feature Loss (DFL)
  - Feed the enhanced waveform to Wav2Vec2 model
  - Compute the log mean-square-error (log-MSE) loss between the last transformer layer's output of enhanced speech and the target speech
  - Only back-propagate the gradient to the enhancement models.

pytorch_stoi: https://github.com/mpariente/pytorch_stoi
Wav2Vec2: https://pytorch.org/audio/stable/pipelines.html#wav2vec2-asr-base-960h

DNN_1 Results

- Complex spectral mapping (DNN_1, DCCRN and Demucs v2 and v3) consistently obtain higher STOI than Conv-TasNet + MVDR.
  - Complex spectral mapping tends to have better alignment estimation.

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**

DNN_1 Results

- DNN_1 significantly outperforms other models without relying on STOI and ASR-based DFL losses.

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**

DNN_1 + mfMCWF Results

- DNN_1 + **single-frame** MCWF degrades the performance on WER and STOI

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

DNN_1 + mfMCWF Results

- Multi-frame MCWF improves DNN_1

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

DNN1 + mfMCWF Results

- Magnitude-mask based mfMCWF*
  underperforms the proposed mfMCWF.

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

*Wang, Zhong-Qiu, Hakan Erdogan, Scott Wisdom, Kevin Wilson, Desh Raj, Shinji Watanabe, Zhuo Chen, and John R. Hershey. "Sequential multi-frame neural beamforming for speech separation and enhancement." In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 905-911. IEEE, 2021.

DNN_1 + mfMCWF + DNN_2 Results

- Adding DNN_2 to DNN_1 + single-frame MCWF improves the performance.

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

DNN_1 + mfMCWF + DNN_2 Results

- Adding DNN_2 to DNN_1 + mfMCWF achieves ~1% improvement.

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

DNN_1 + mfMCWF + DNN_2 Results

- DNN_1 + two iterations of (mfMCWF + DNN_2) achieves the best performance.

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

Conclusions

- We proposed iNeuBe framework, an iterative pipeline of linear beamforming and DNN-based complex spectral mapping.
- Computing mfMCWF weights using DNN-based complex spectral mapping output can have significant advantages in the challenge scenario.
- Comparing with multiple state-of-the-art models, iNeuBe framework achieves remarkably better challenge metrics, with both lower WER and higher STOI, even when the competing models are trained with back-end task aware losses.