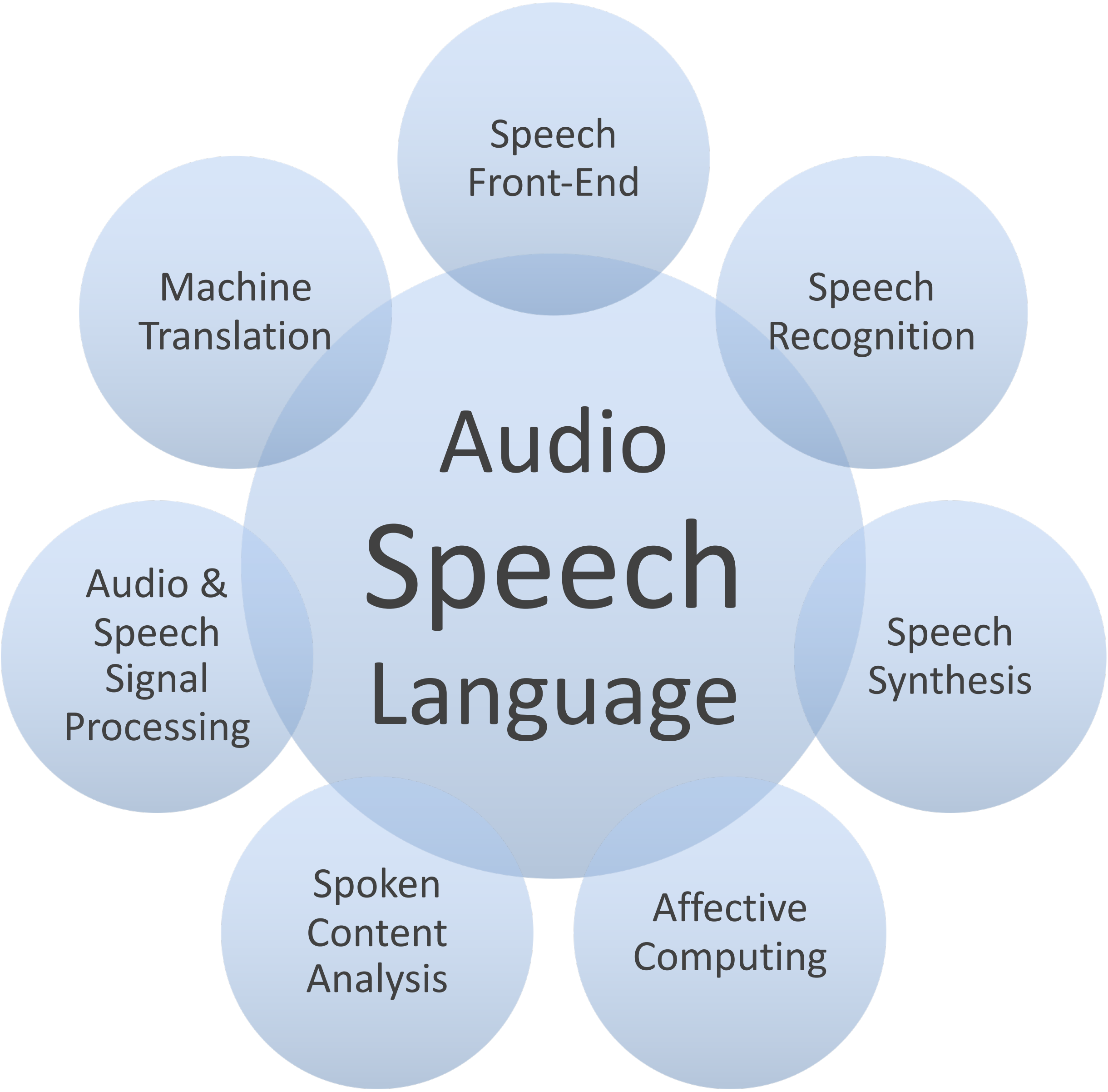
Recent Advances in Inaudible Adversarial Attack in Speaker Recognition and Multi-channel Speech Separation in Complicated Environments

Lei Xie
Audio, Speech & Language Processing Group (ASLP@NPU),
Northwestern Polytechnical University, Xi'an, China

音频语音与语言处理研究组
Audio, Speech and Language Processing Group NPU

ASLP@NPU


Speech Front-End
Machine Translation
Speech Recognition
Audio
Speech
Language
Audio & Speech Signal Processing
Speech Synthesis
Spoken Content Analysis
Affective Computing

Outline

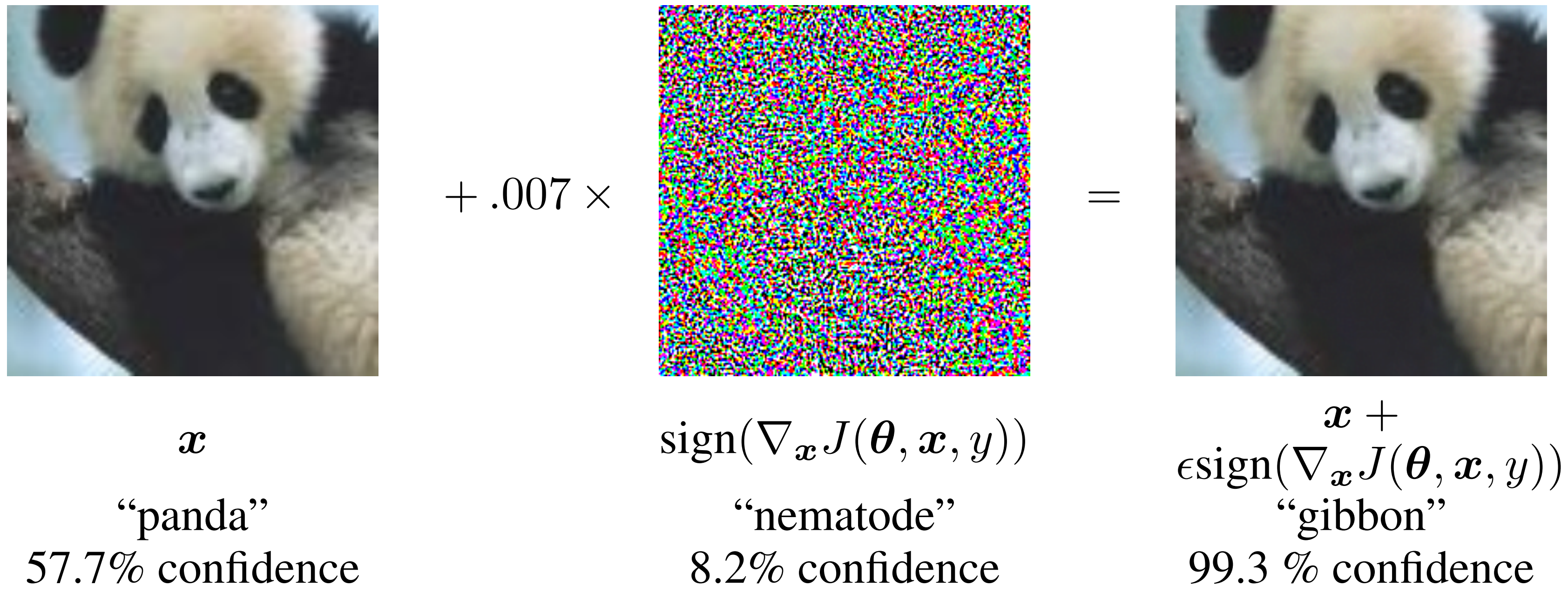
- Inaudible Adversarial Attack in Speaker Recognition
- Multi-channel Speech Separation in Complicated Environments

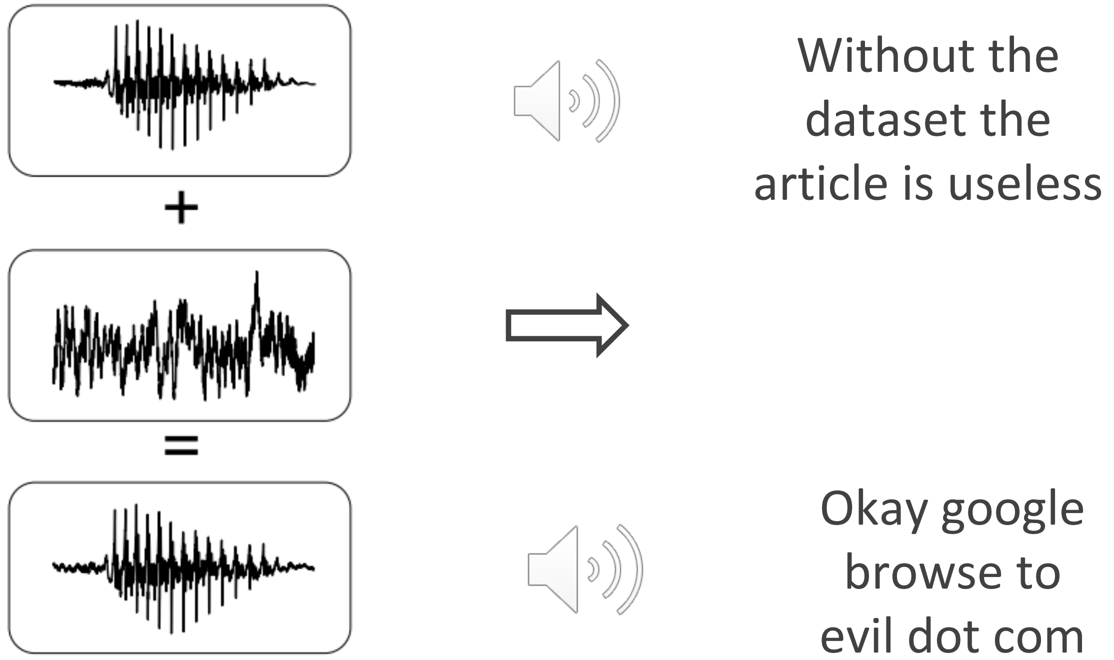
Outline

- **Inaudible Adversarial Attack in Speaker Recognition**
- Multi-channel Speech Separation in Complicated Environments

Adversarial Attacks in Speaker Recognition

- Spoofing attacks: reply, TTS, VC, etc
- DNNs are also vulnerable to adversarial examples (e.g. image or speech related tasks)
- **Adversarial examples**:
  - Examples with small, intentional perturbations that cause a well-trained model make a false prediction





Figures and samples are from Goodfellow 2014 [1] and Carlini 2018 [2].

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.
[2] N. Carlini and D. Wagner, "Audio adversarial examples: targeted attacks on speech-to-text," in Security and Privacy Workshops (SPW). IEEE, 2018, pp. 1–7.

Adversarial Attacks in Speaker Recognition

- **Performing Adversarial Attacks**
  - **Goal of the attacker**
    - **Adversarial impersonation** → targeted attack (user authentication application)
      - Attack transforms a non-target trail (random spkr) into a target trail (target spkr)
      - Attacker wants to usurp the identity of another person
    - **Adversarial evasion** → non-targeted attack (forensics, criminal investigation)
      - Attacks transform a target-trail (target spkr) into non-target (different spkr)
      - Attacker wants to avoid detection by ASV system
  - **Knowledge of the attacker**
    - White-box: has full knowledge of the system under attack
    - Black-box: has no access to the victim model, generates adv. speech using another white-box system
    - Grey-box: has some information, but not statistical models
  - Methods of the generation of adversarial examples: FGSM, iterative FGSM, Carlini-Wagner...

[3] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, “Fooling end-to-end speaker verification with adversarial examples,” in IEEE ICASSP 2018, 2018, pp. 1962–1966.
[4] G. Chen, S. Chen, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, “Who is real Bob? adversarial attacks on speaker recognition systems,” ArXiv, vol. abs/1911.01840, 2019.
[5] Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, “Practical adversarial attacks against speaker recognition systems,” in ACM HotMobile 2020, 2020, pp. 9–14.
[6] Das, R.K., Tian, X., Kinnunen, T. and Li, H., 2020. The Attacker's Perspective on Automatic Speaker Verification: An Overview. in Interspeech 2020, pp.4213-4217.
[7] Villalba, J., Zhang, Y. and Dehak, N., 2020. x-Vectors Meet Adversarial Attacks: Benchmarking Adversarial Robustness in Speaker Verification. in Interspeech 2020, pp.4233-4237.
[8] Zhang, Y., Jiang, Z., Villalba, J. and Dehak, N., 2020. Black-box Attacks on Spoofing Countermeasures Using Transferability of Adversarial Examples. Proc. Interspeech 2020, pp.4238-4242.

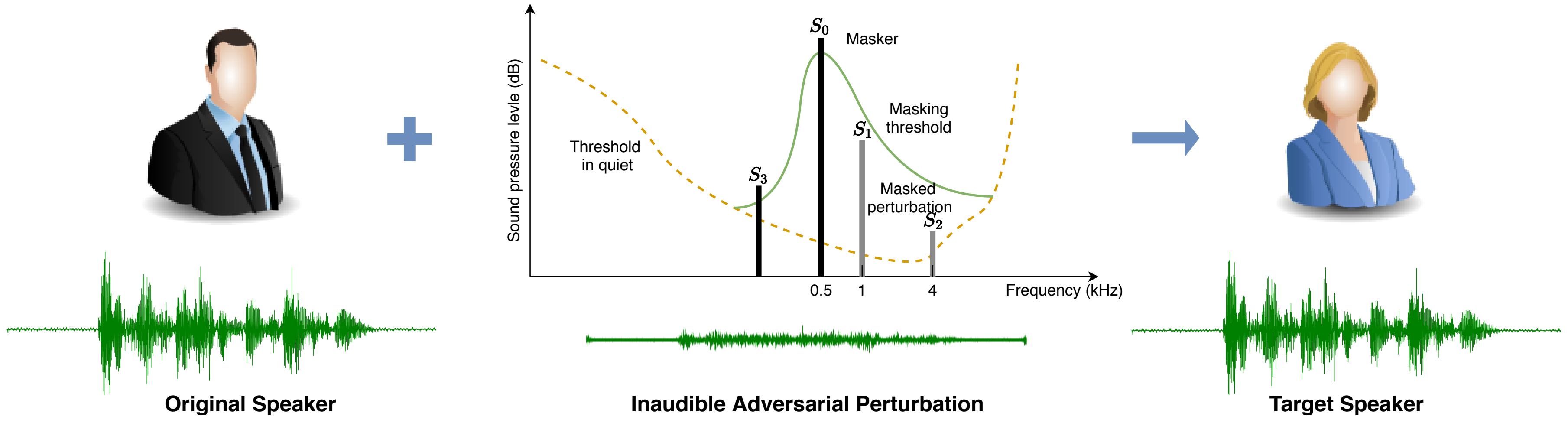
Adversarial Attacks in Speaker Recognition

- **Defenses of adversarial attacks**
  - **Improve the robustness of SV model against adversarial attacks**
    - Adversarial regularization is proposed to protect end-to-end speaker verification system [9]. This mechanism aims at finding a worst spot around the current data point, and then optimize using this worst data point to derive a robust model.
  - **Defense against adversarial attacks**
    - A passive defense method--spatial smoothing and another proactive method--adversarial training are studied to defend adversarial attacks for spoofing countermeasures [10].
  - **Detection of adversarial examples**
    - Defend ASV systems against adversarial attacks with a separate detection network [11]. A VGG-like binary classification detector is introduced and demonstrated to be effective on detecting adversarial samples.

[9] Q. Wang, P. Guo, S. Sun, L. Xie, and J. H. Hansen, “Adversarial regularization for end-to-end robust speaker verification,” in Interspeech 2019, 2019, pp. 4010–4014.
[10] H. Wu, S. Liu, H. Meng, and H. yi Lee, “Defense against adversarial attacks on spoofing countermeasures of ASV,” in IEEE ICASSP 2020, 2020, pp. 6564–6568.
[11] Li, X., Li, N., Zhong, J., Wu, X., Liu, X., Su, D., Yu, D. and Meng, H., 2020. Investigating Robustness of Adversarial Samples Detection for Automatic Speaker Verification. in Interspeech 2020, pp.4233-4237.

Inaudible Adversarial Perturbations for Targeted Attack in Speaker Recognition

- In our study, we aim to exploit this weakness to perform targeted adversarial attacks against speaker recognition system
- The aforementioned adversarial examples are mostly restricted to make a slight change of original signal in form of audio sampling points, without considering the human sound perceptibility
- Our aim: Generate inaudible adversarial perturbations for targeted attacking speaker recognition system on wave-level.
- Our approach: Leverage frequency masking [12]
  - **Audible sound (random speaker) + another louder audible sound (perturbation) → inaudible sound (inaudible adv. example)**
- Explore the targeted attacks on non-speech



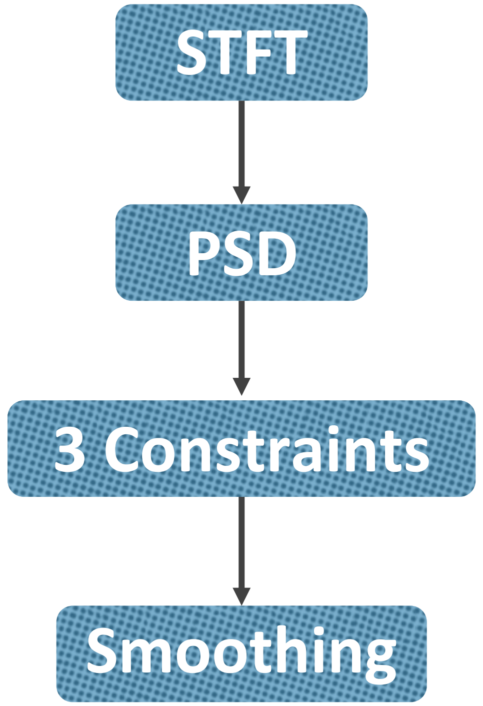
An overview of the generation of adversarial examples based on frequency masking.

- **Cons of previous adversarial perturbations:**
  - Perturbations are small → lower attack success rate
  - Constrict noise by l_p norm → easily detectable

[12] Qing Wang, Pengcheng Guo, Lei Xie, Inaudible Adversarial Perturbations for Targeted Attack in Speaker Recognition, Interspeech2020 https://arxiv.org/abs/2005.10637

Inaudible Adversarial Perturbations for Targeted Attack in Speaker Recognition

- **Estimation of frequency masking threshold**
  - **Step1**: Identifications of maskers



$$P_x(k)/dB = 10\log_{10}\left|\frac{1}{N}s_x(k)\right|^2$$

a) Local maxima;

b) Larger than absolute threshold of hearing

c) Maximum amplitude within 0.5 Bark

$$\bar{P}_x(\bar{k}) = 10\log_{10}\left[10^{\frac{\bar{P}_x(k-1)}{10}} + 10^{\frac{\bar{P}_x(k)}{10}} + 10^{\frac{\bar{P}_x(k+1)}{10}}\right]$$

- **Step2**: Calculation of individual masking thresholds
  - $T[b(j), b(i)]$: masker at j-th freq. contributes to the masking threshold on maskee at i-th freq.

$$T[b(j), b(i)]/dB = \bar{P}_x[b(j)] + \triangle\,[b(j)] + SF[b(j), b(i)]$$

- **Step3**: Calculation of global masking threshold

$$T_G(i)/dB = 10\log_{10}\left[10^{\frac{ATH(i)}{10}} + \sum_{j=1}^{N_M} 10^{\frac{[b(j),b(i)]}{10}}\right]$$

[13] Y. Lin and W. H. Abdulla, "Principles of psychoacoustics," in Audio Watermark. Springer, 2015, pp. 15–49.

Inaudible Adversarial Perturbations for Targeted Attack in Speaker Recognition

- **Objective functions**

$$L_{TH}(x,\delta) = \mathbb{E}_k max\{\bar{P}_\delta(k) - T_G(k), 0\}$$

$$\min L(x,\delta,y') = L_{CE}(f(x+\delta),y') + \alpha \cdot L_{TH}(x,\delta)$$

- **Optimization procedure**

Attack Stage 1:

$$\delta \leftarrow clip_\epsilon\left(\delta - lr_1 \cdot sign\big(\nabla_\delta L_{CE}(f(x+\delta),y')\big)\right)$$

Attack Stage 2:

$$\delta \leftarrow \delta - lr_2 \cdot \nabla_\delta L(x,\delta,y')$$

- **Cons of previous adversarial perturbations:**
  - Perturbations are small → lower attack success rate
  - Constrict noise by l_p norm → easily detectable

- **Pros of inaudible adversarial perturbations:**
  - Perturbations can be larger and inaudible
  - Constrict function is consistent with psychoacoustic principle

Inaudible Adversarial Perturbations for Targeted Attack in Speaker Recognition

- **Dataset**
  - Aishell-1:
    - Original set: 10 female (F) and 10 male speakers (M), each with 100 utterances
    - Attack target set: another 10 female (F’) and 10 male speaker (M’), each with 100 utterances
    - Four test modes: M2M’, M2F’, F2M’ and F2F’
  - MUSAN (Music portion from MUSAN as the non-speech dataset):
    - 200 pieces of western art music are cut into 1000 pieces of 6 seconds short segments
  - Room Impulse Response and Noise Database
    - Used for on-the-air attack

- Baseline
  - White-box attack: x-vector system [13]
  - On-the-air attack: SincNet system [14]

- Evaluation metric
  - Attack success rate

    $$Acc = N_s \,/\, N$$

  - Perceptual evaluation of speech quality (PESQ)
  - Signal-to-noise ratio (SNR)
  - Subjective listening

[14] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S., 2018, April. X-vectors: Robust dnn embeddings for speaker recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5329-5333). IEEE.
[15] Ravanelli, M. and Bengio, Y., 2018, December. Speaker recognition from raw waveform with sincnet. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 1021-1028). IEEE.

Inaudible Adversarial Perturbations for Targeted Attack in Speaker Recognition

- Experimental results and analysis
  - **White-box attack (x-vector system)**

System	M2M'	M2F'	F2M'	F2F'
Attack Stage1	72.6	73.8	73.3	71.3
Attack Stage2	98.5	97.6	96.7	93.8
Before Attack				
Attack Stage1				
Attack Stage2				

Audio samples

Perturbation

  - **On-the-air attack (SincNet system)**

System	M2M'	M2F'	F2M'	F2F'
Attack Stage1	4.8	3.9	4.5	3.7
Attack Stage2	47.1	45.2	42.1	41.6
Before Attack				
Attack Stage1				
Attack Stage2				

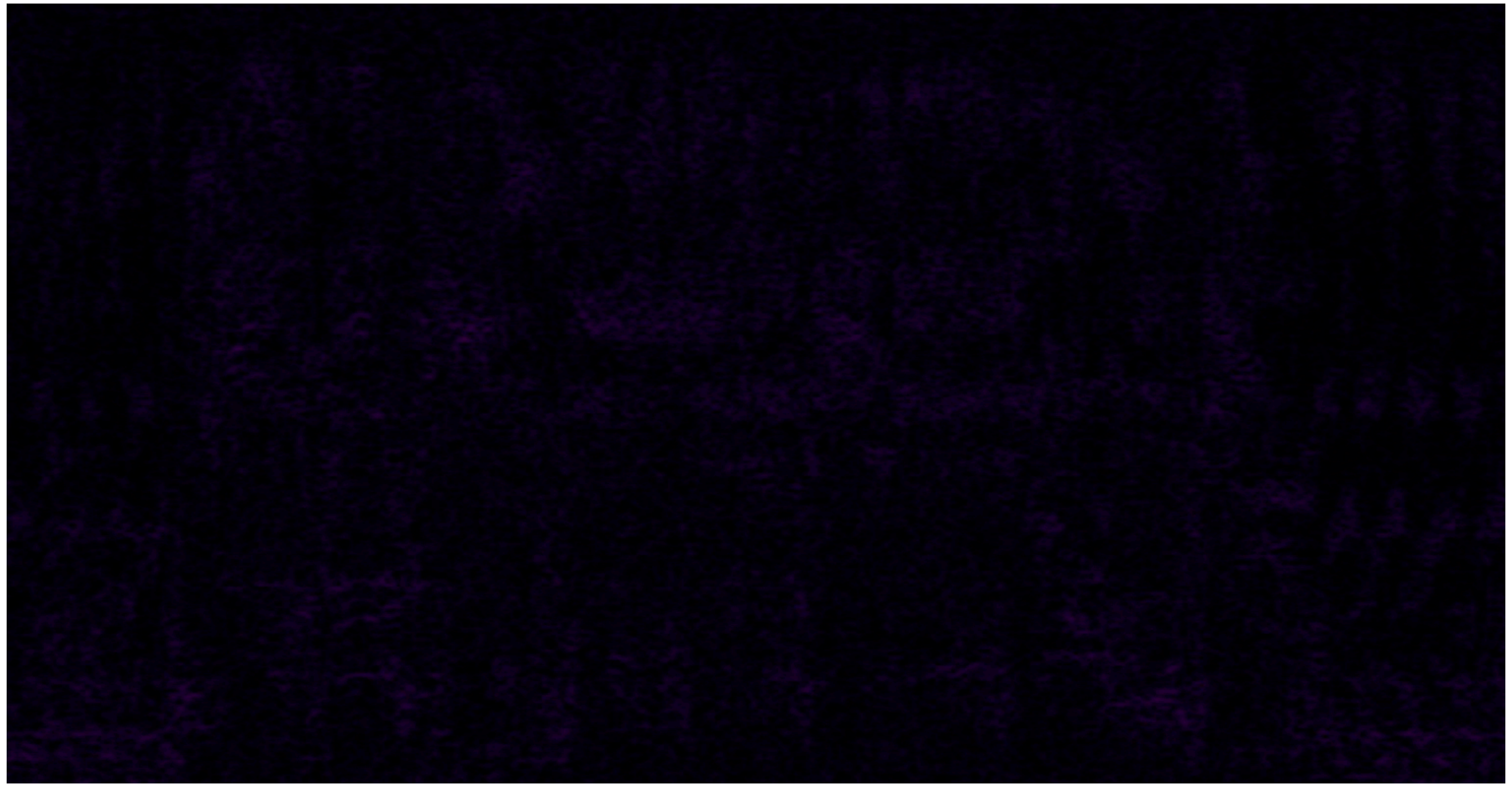
Audio samples

- White-box attack yields up to 98.5% attack success rate to arbitrary gender speaker targets with inaudible adversarial perturbations
- We can achieve up to 47.1% attack success rate in on-the-air attack

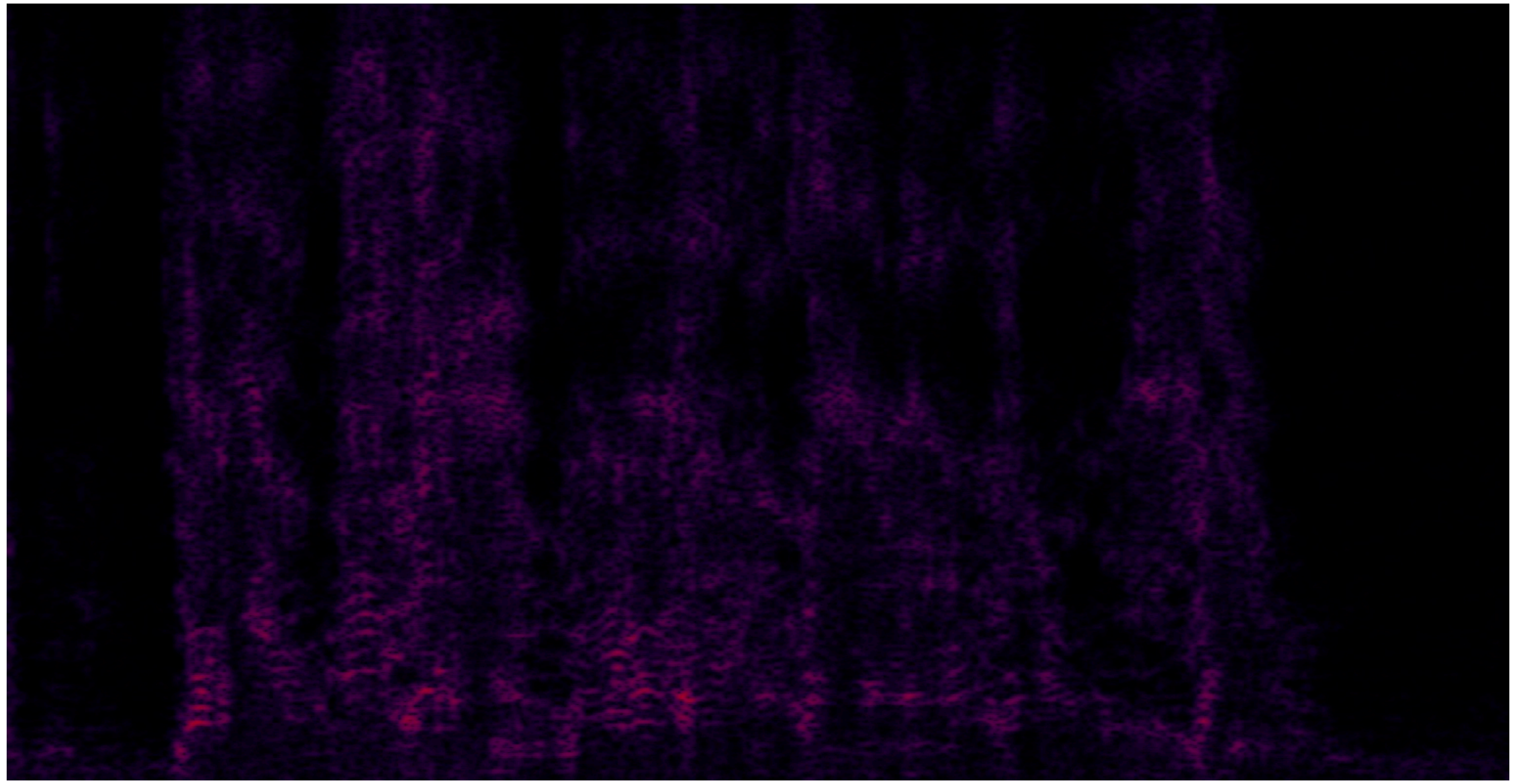
More demos: https://pengchengguo.github.io/inaudible-advex-for-sv/

Inaudible Adversarial Perturbations for Targeted Attack in Speaker Recognition

- Larger perturbation by the proposed approach



M2M’ -- Stage 1



M2M’ -- Stage 2

More demos: https://pengchengguo.github.io/inaudible-advex-for-sv/

Inaudible Adversarial Perturbations for Targeted Attack in Speaker Recognition

- Experimental results
  - **White-box attack**
    - Objective evaluation

PESQ and SNR (dB) comparison of Attack Stage1 and Attack Stage2.

- **Subjective listener evaluation**

Preference (%)			*p*-value
Attack Stage1	Neural	Attack Stage2	
11.33	20.00	**68.67**	0.0379

Preference scores (%) of Attack Stage1 and Attack Stage2.

- **Non-speech targeted attack**

	Before Attack	Attack Stage1	Attack Stage2
Acc	0.00%	77.0%	91.5%
Sample1			
Sample2			

- Conclusions
  - Objective and subjective evaluations indicate that frequency masking based adversarial perturbations are more inaudible, even with larger absolute energies
  - Experiments on MUSAN corpus show that even non-speech can achieve a high targeted speaker attack success rate.

[12] Qing Wang, Pengcheng Guo, Lei Xie, Inaudible Adversarial Perturbations for Targeted Attack in Speaker Recognition, Interspeech2020 https://arxiv.org/abs/2005.10637

Future Directions and Challenges

- More realistic scenarios:
  - On-the-air attack
  - Black-box attack
- Defense/detection of adversarial attacks
- Also some other challenges:
  - Evaluation metrics?
  - Standard dataset?
  - Any other attack scenario?

Outline

- Inaudible Adversarial Attack in Speaker Recognition
- **Multi-channel Speech Separation in Complicated Environments**

Move to the Cocktail Party Problem

"One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. This is such a common experience that we may take it for granted; we may call it 'the cocktail party problem'..." (Cherry' 57)

Towards Multi-Talker Speech Recognition

- Speech separation is a common practice to handle the speaker overlaps
- **Multi-talker aware ASR**
  - MIMO-Speech, SpeakerBeam…
- **Front-end + Back-end**
  - Beamforming, esp. Fixed Beamforming
  - Mask-based Adaptive Beamforming
  - Ad-hoc Speech Enhancement and Separation
- Speaker-independent Continuous Speech Separation (SI-CSS)
- Injecting prior knowledge (bias) into speech separation

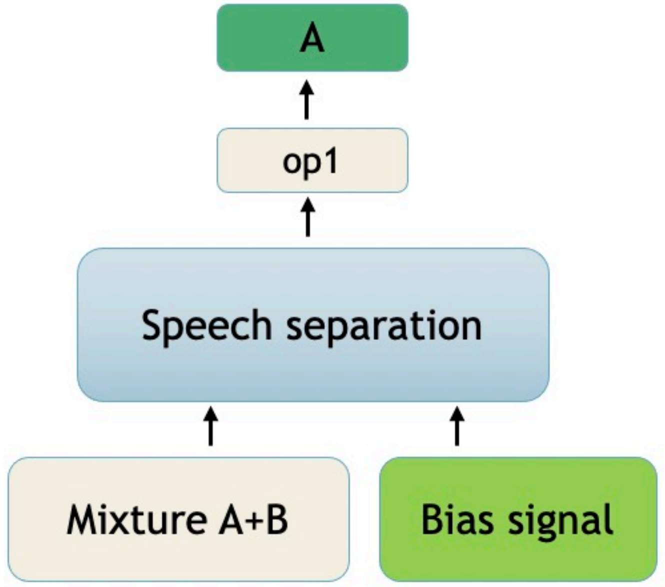
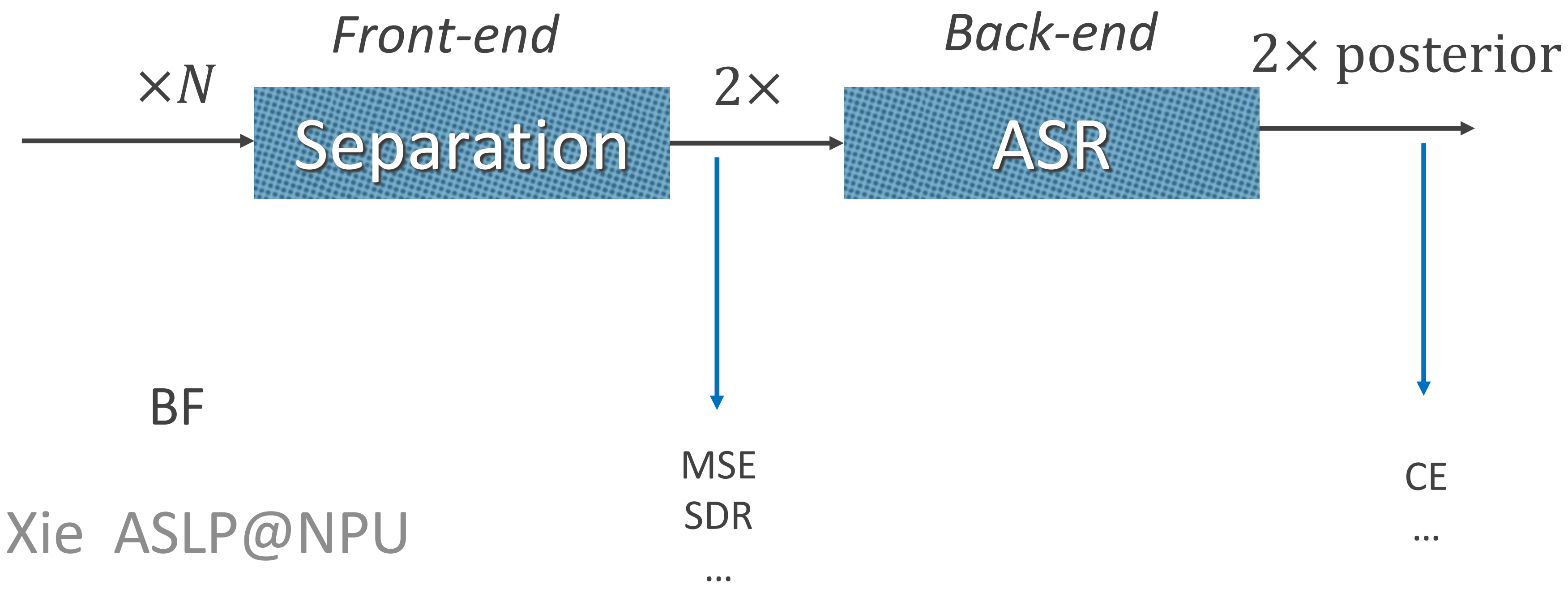
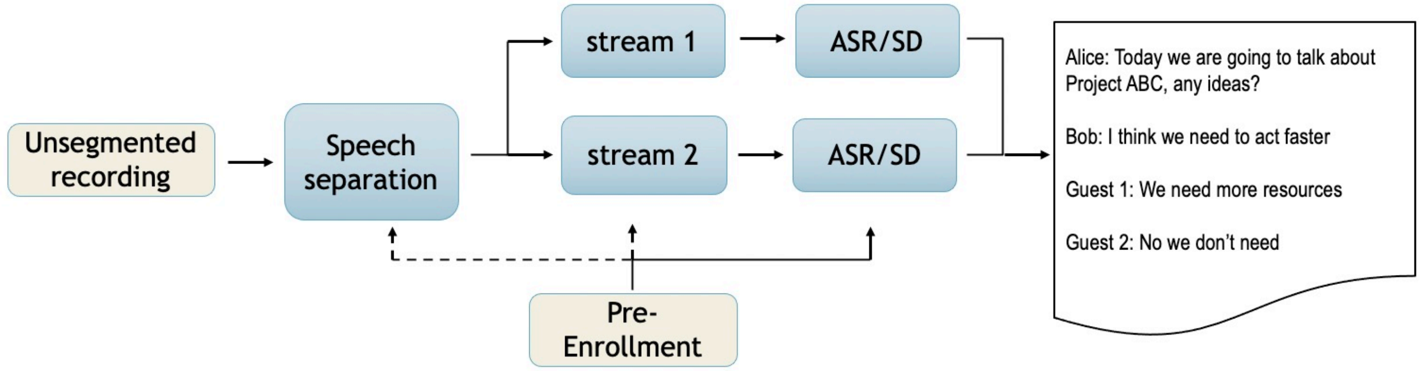
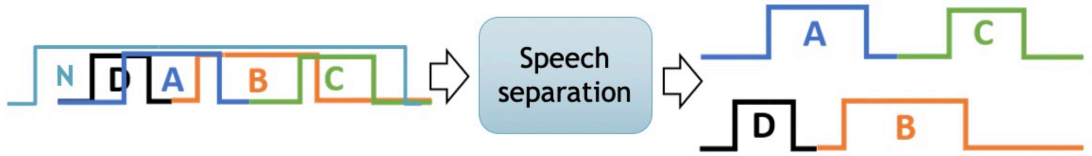


Figure Credit: Zhuo Chen







Low-latency Continuous Speech Separation

- Extraction vs. Separation
  - Speech extraction usually has better performance upper bound and is easier to joint train with other module
  - But it usually suffers from the efficiency limitation and heavily depend on the bias signal
- **UFE**: Combining the advantageous from both [1]
  - Speech separation pre-separate the mixed signal
  - Speech extraction further enhance the result
  - Acceptable computation cost with low latency online processing

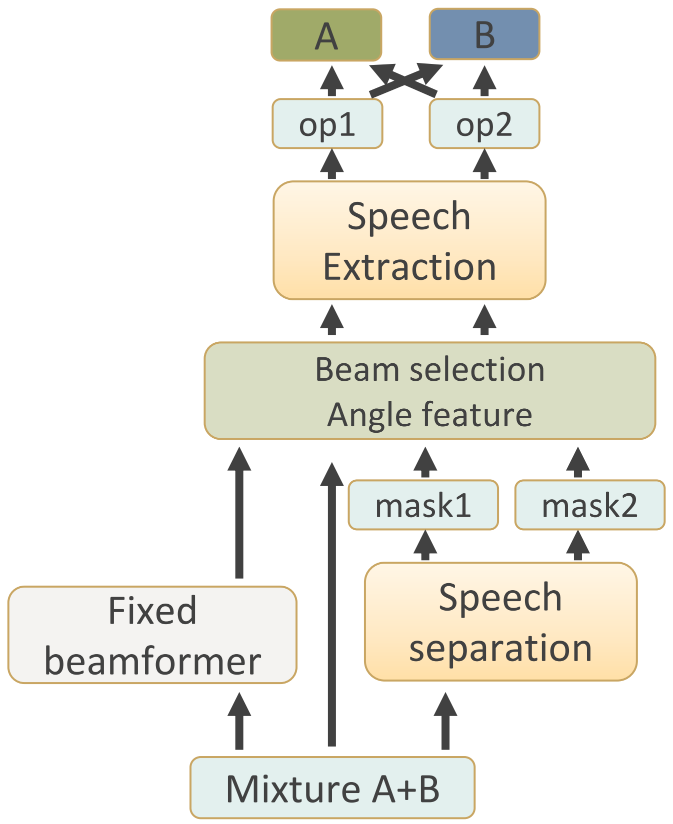
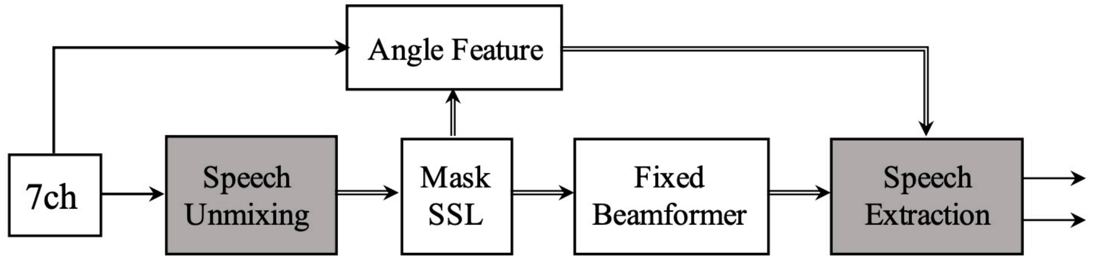


Figure Credit: Zhuo Chen

[1] Takuya Yoshioka, Zhuo Chen, Changliang Liu, Xiong Xiao, Hakan Erdogan, and Dimitrios Dimitriadis, "Low-latency speaker-independent continuous speech separation," ICASSP 2019

E2E Online Multi-Channel Speech Separation



- UFE System (**U**nmixing, **F**ixed-beam and **E**xtraction) [1]
  - *M*-channel STFT of input speech mixture: $Y_{0,\ldots,M-1} = \{Y_0, \ldots, Y_{M-1}\}$
  - **Unmixing network (U)**: multi-channel TF mask $\mathbf{M}_{0,1} \in \mathbb{R}^{T\times F}$ estimation via PIT under Si-SNR loss

$$\mathcal{L} = -\max_{\phi\in\mathcal{P}} \sum_{(i,j)\in\phi} \text{Si-SNR}(\mathbf{s}_i, \mathbf{x}_j), \qquad \mathbf{s}_i = \text{iSTFT}\left(\mathbf{M}_i \odot \mathbf{Y}_0\right)$$

  - Sound Source Localization (SSL): estimate the spatial angle for i^{th} speaker

$$\mathcal{D}_{\theta,i} = -\sum_{t,f} \mathbf{M}_{i,tf} \log\left(1 - \frac{|\mathbf{y}_{t,f}^H \mathbf{h}_{\theta,f}|^2}{1+\epsilon}\right)$$

  - **Fixed beamformer (F)**

$$b_{i,t,f} = \mathbf{w}_{i,f}^H \mathbf{y}_{t,f}, \quad \mathbf{y}_{t,f} = [\mathbf{Y}_{0,tf}, \cdots, \mathbf{Y}_{M-1,tf}]^T$$

  - **Extraction Network (E):** location-based speech extraction on each selected beam

$$\mathbf{a}_{\theta,f} = \frac{1}{P} \sum_{i,j\in\psi} \cos(\mathbf{o}_{ij,f} - \Delta_{\theta,ij,f}),$$

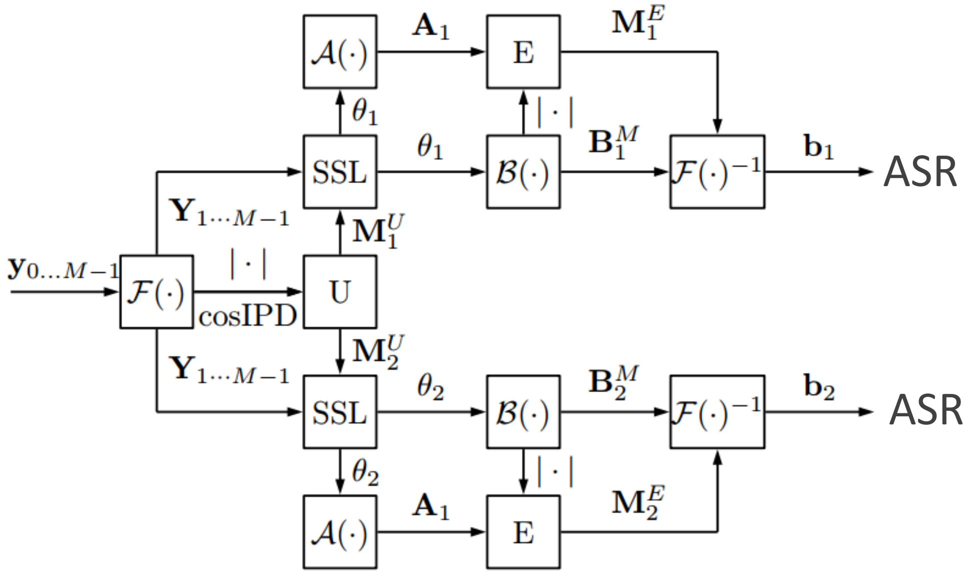


Fig. 1. Overview of the UFE system. $\mathcal{F}$, $\mathcal{B}$, $\mathcal{A}$ and SSL denote short-time Fourier transform (STFT), fixed beamforming, angle feature computation and SSL algorithm, respectively. $\mathbf{M}_i^U$ and $\mathbf{M}_i^E$ represent the TF-masks of the i-th speaker generated by *unmixing* (U) and *extraction* (E) network. $\mathbf{A}_i$ and $\mathbf{B}_i^M$ denote the angle feature and the selected beam given the speaker direction θ_i. The *unmixing* and *extraction* model are trained independently.

[1] Takuya Yoshioka, Zhuo Chen, Changliang Liu, Xiong Xiao, Hakan Erdogan, and Dimitrios Dimitriadis, "Low-latency speaker-independent continuous speech separation," ICASSP 2019

E2E Online Multi-Channel Speech Separation

- Advantages of the UFE system
  - Low latency as fixed beamformer used
  - Overcome the weak spatial cancellation issue for common fixed beamformer applications through additional speech extraction step
- Drawbacks of the UFE system: **modularized optimization** with sub-optimal performance
  - All components are optimized separately
    - Speech unmixing and extraction are optimized with **signal reconstruction metric**
    - Sound localization is optimized with **ML**
    - Beamformer is designed with **hand tuned criteria**

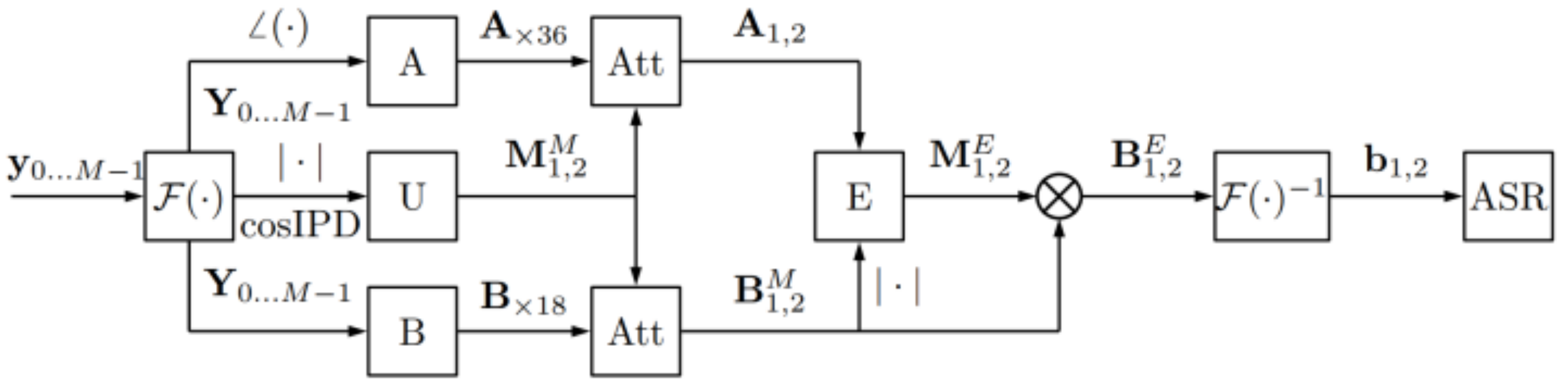
E2E Online Multi-Channel Speech Separation

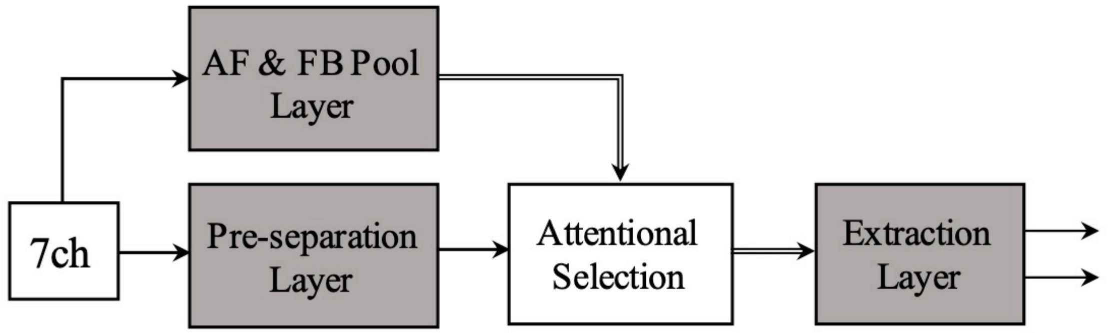
- Proposed E2E-UFE: all components are optimized jointly via a unified network [2]
  - The TF-masks generated by unmixing network is converted to **hidden representation**
  - **An attentional module** between the mask-embedding and beamforming output, candidate directional features is applied to pick the corresponding beam and angle feature, which are passed to neural extraction module
    - Allow the gradients to propagate though the beam selection module, which was non-differentiable in the original UFE
  - Extraction network takes both beams and angle features as input, **outputting two beams simultaneously**
  - All the outputting beams are **optimized jointly with PIT objective**, which avoids the permutation ambiguity when speakers are spatially close.
- With these updates, we ensure that the gradient from the top layer can pass to all sub-modules of the system, i.e. making the system **optimized in an end-to-end manner**, while keep the advantage of base model with low-latency processing

[2] Jian Wu, Zhuo Chen,Jinyu Li,Takuya Yoshioka, Zhili Tan, Ed Lin, Yi Luo, Lei Xie, AN END-TO-END ARCHITECTURE OF ONLINE MULTI-CHANNEL SPEECH SEPARATION, Interspeech2020 https://arxiv.org/abs/2009.03141

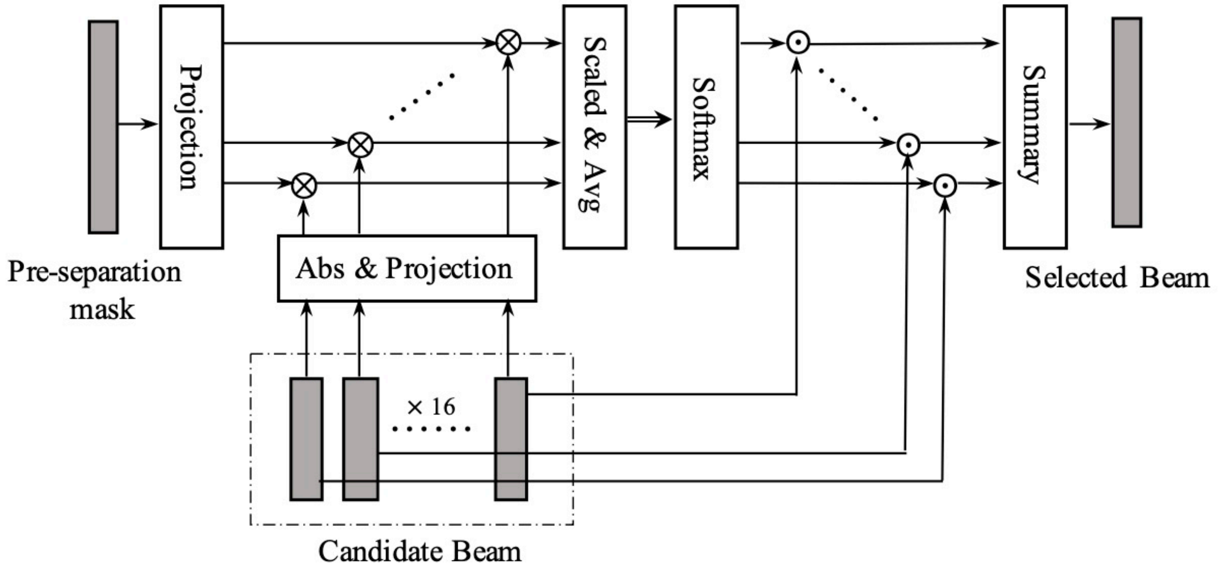
E2E Online Multi-Channel Speech Separation

- Proposed E2E-UFE





- Pre-separation (U)
- Attentional beam & angle selection (F)



$$\mathbf{V}^P = \mathbf{E}\mathbf{W}_p,$$
$$\mathbf{V}^B = |\mathbf{B}|\mathbf{W}_b,$$

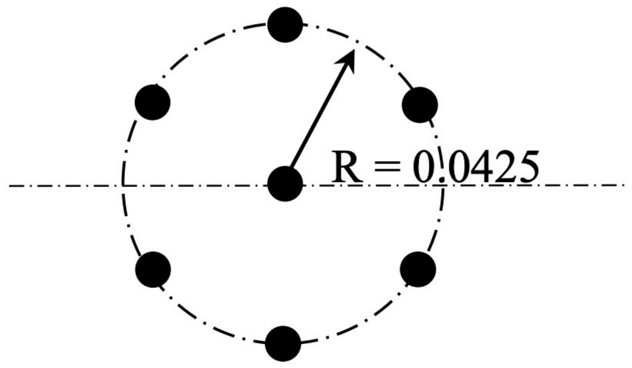
$$s_{h,b,t} = (\sqrt{D})^{-1} \left(\mathbf{V}^P_{h,t}\right)^T \mathbf{V}^B_{b,t}$$
$$\hat{s}_{h,b} = (T)^{-1} \sum_t s_{h,b,t},$$
$$w_{h,b} = \text{softmax}_b(\hat{s}_{h,b}),$$
$$\hat{\mathbf{B}}_h = \sum_b w_{h,b}\mathbf{B}_b.$$

- Joint Extraction (E): Beam selection & wave reconstruction are optimized with one objective function

$$\mathcal{L} = -\max_{\phi \in \mathcal{P}} \sum_{(i,j)\in\phi} \text{Si-SNR}(\mathbf{s}_i, \mathbf{r}_j),$$

E2E Online Multi-Channel Speech Separation

- Experiments
  - Training data
    - On-the-fly data simulation using Librispeech + three Microsoft's internal dataset
    - Additional isotropic noise is used
    - Overlapping ratio: 0.5 ~ 1.0
    - Speaker angle: at least 20 degrees
    - Distance between speaker and array: at least 1m
  - Evaluation data
    - Two dataset: *simu* and *semi-real*
    - *simu* - simulated with *dev* set in Librispeech
    - *semi-real* - simulated with real recordings
    - Two overlapping ratio: 0.2~0.5 (OV35) & 0.5~1.0 (OV75)
- Feature
  - STFT: 32/16ms
  - cosIPD pair: 1,4/2,5/3,6
  - Angle feature: 1,0/2,0/3,0/4,0/5,0/6,0
- Network Configurations
  - U & E: 3 Contextual LSTM layers with 512 nodes
  - Add future context for uni-directional LSTMs
- Evaluation metric
  - WER
  - Offline & Online



E2E Online Multi-Channel Speech Separation

- Results

Offline evaluation

Method	*simu* OV35	*simu* OV75	*semi-real* OV35	*semi-real* OV75
Mixed Beam	67.40	52.40	70.92	57.63
Clean Beam	10.67	10.56	20.34	19.71
UFE	16.44	18.55	35.60	37.54
E2E-UFE	**16.85**	**18.98**	**33.89**	**35.92**

Block online evaluation

Method (history)	*simu* OV35	*simu* OV70	*semi-real* OV35	*semi-real* OV70
UFE (2s)	24.10	31.40	44.05	45.13
UFE (4s)	23.66	28.85	43.49	44.06
E2E-UFE (2s)	17.50	19.43	38.64	39.98
E2E-UFE (4s)	**17.09**	**19.10**	**36.67**	**39.11**

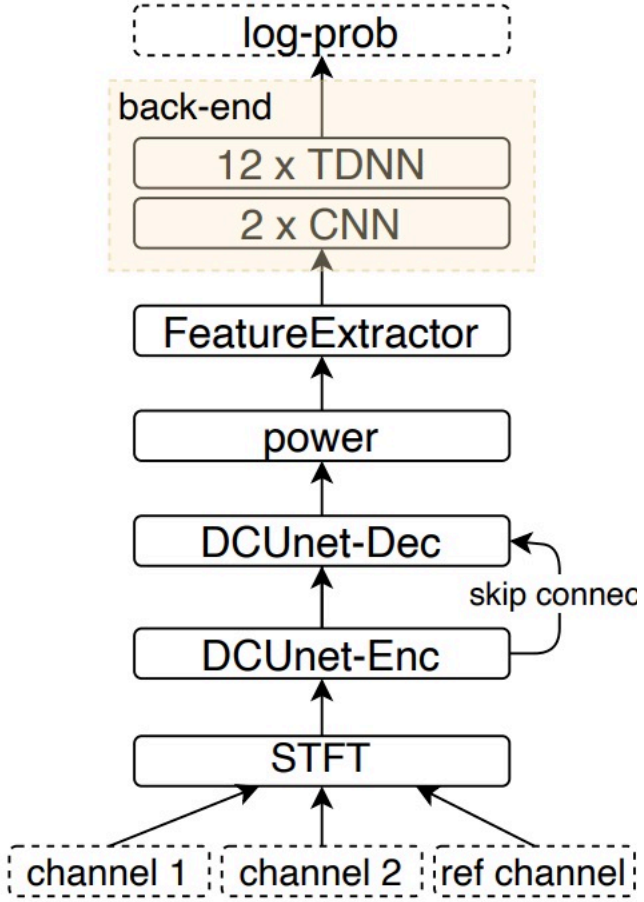
- Simple FB (Mixed Beam) yielded a high WER even with oracle DoA while Clean Beam sets the upper bound
- The proposed E2E-UFE achieved comparable performance as the original UFE for the simulated data set, while demonstrating a clear performance advantage in *semi-real set*
- E2E- UFE shows robustness for different look-back configurations (a 2s or 4s history context), achieving slightly worse results than for the offline evaluation on both datasets
- Original UFE resulted in a much larger performance degradation for the online evaluation
- On the *semi-real* set, E2E-UPE brought about a 12.47% average relative WER reduction compared with UFE using a 2 s history context, while on the *simu* set, the relative reduction increases to 29.71%

[2] Jian Wu, Zhuo Chen,Jinyu Li,Takuya Yoshioka, Zhili Tan, Ed Lin, Yi Luo, Lei Xie, AN END-TO-END ARCHITECTURE OF ONLINE MULTI-CHANNEL SPEECH SEPARATION, Interspeech2020 https://arxiv.org/abs/2009.03141

DCUNET Front-end for Multi-channel ASR

- Adopt the architecture of deep complex Unet (**DCUnet**) - a powerful **complex-valued Unet-structured** speech enhancement model - as the front-end of multi-channel acoustic model
- Integrate them in a multi-task learning (MTL) framework along with cascaded framework
  - **DCUnet-MTL**
  - **DCUnet-CAS**
- Experiments: 1000-hours real-world XiaoMi smart speaker data with echoes
  - DCUnet-MTL method brings 12.2% relative CER reduction compared with the traditional approach with array processing + single-channel acoustic model
  - It also achieves superior performance over the recently proposed neural beamforming method

Does a strong enhancement network benefit ASR?



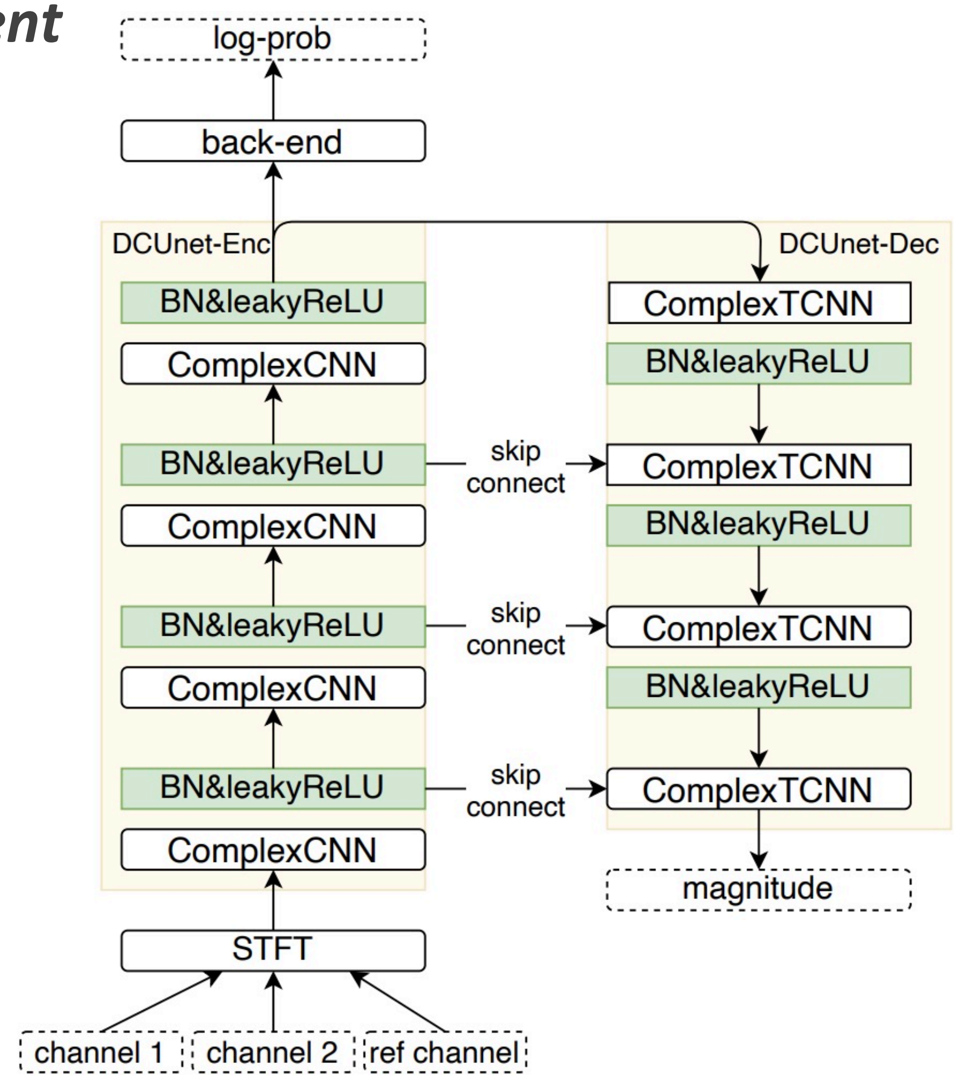


Table 4. CER (%) comparison on different subsets

Model	Echoed	<5 dB	[5,15) dB	⩾15 dB	Total
Baseline	16.49	19.67	14.16	12.36	15.08
NNFB	15.80	18.61	13.82	12.45	14.67
DCUnet-MTL	**14.68**	**16.11**	**12.54**	**11.18**	**13.23**
DCUnet-CAS	15.11	17.55	12.91	11.64	13.82

[3] Yuxiang Kong, Jian Wu, Quandong Wang, Peng Gao, Weiji Zhuang, Yujun Wang, Lei Xie, Multi-Channel Automatic Speech Recognition Using Deep Complex Unet, IEEE SLT2021, https://arxiv.org/abs/2011.09081

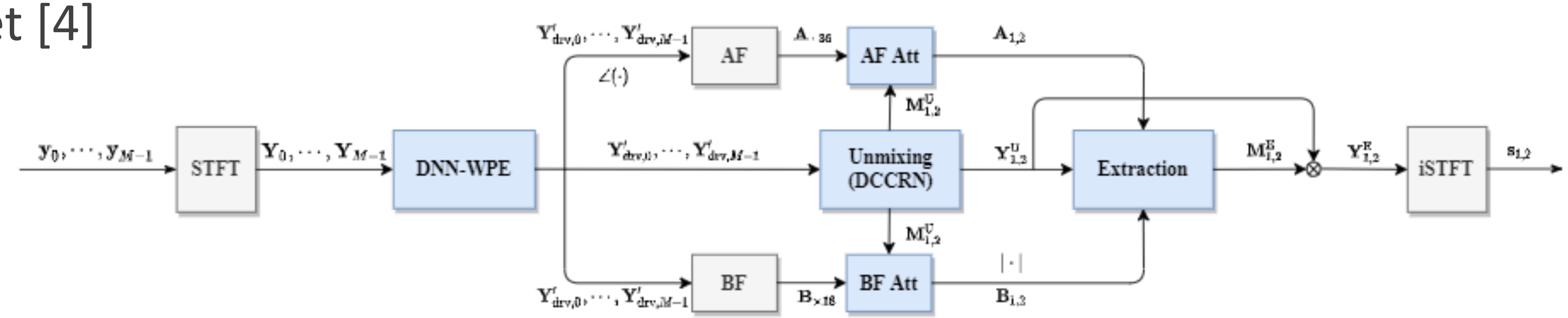
DESNet: A Multi-Channel Network for Simultaneous Speech Dereverberation, Enhancement and Separation

- Motivation
  - Real-world environment: speech overlapping, directional/isotropic noise and reverberation may exist together
  - Prior arts: Direct separation on noisy mixtures, cascaded/two-stage (enhancement-separation, separation-enhancement), recursive separation…
  - **E2E-UFE**[2] and **DCCRN**[4] show great potential on multi-channel separation and single-channel enhancement
- Contribution
  - We propose an offline processing neural network for simultaneous speech **D**ereverberation, **E**nhancement and **S**eparation (**DESNet**)
  - We combine the **DNN-WPE, E2E-UFE and DCCRN** organically together with differentiable STFT (iSTFT) to form an end-to-end manner
- We evaluate the performance of the proposed model
  - Three scenarios: speech enhancement (SE), clean speech separation (CSS) and noisy speech separation (NSS)
  - Two categories: dereverberated and non-dereverberated

[2] Jian Wu, Zhuo Chen,Jinyu Li,Takuya Yoshioka, Zhili Tan, Ed Lin, Yi Luo, Lei Xie, AN END-TO-END ARCHITECTURE OF ONLINE MULTI-CHANNEL SPEECH SEPARATION, Interspeech2020 https://arxiv.org/abs/2009.03141
[4] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, Lei Xie, DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement, Interspeech2020, October 25-29, Shanghai, China https://arxiv.org/abs/2008.00264

DESNet: A Multi-Channel Network for Simultaneous Speech Dereverberation, Enhancement and Separation

- Proposed DESNet [4]



- Dereverberation: DNN-WPE [5]

$$\begin{aligned} \mathbf{y}'_{\mathrm{drv},t,f} &= \mathbf{y}_{t,f} - \sum_{k=0}^{K-1} \mathbf{G}_{f,k}^{H} \mathbf{y}_{t-\Delta-k,f} \\ &= \mathbf{y}_{t,f} - \mathbf{G}_{f}^{H} \overline{\mathbf{y}_{t-\Delta,f}}, \end{aligned}$$

$$\mathbf{R}_f = \sum_t \frac{\overline{\mathbf{y}_{t-\Delta,f}}\,\overline{\mathbf{y}_{t-\Delta,f}}^{H}}{\Lambda_{t,f}},$$

$$\mathbf{r}_f = \sum_t \frac{\overline{\mathbf{y}_{t-\Delta,f}}\,\mathbf{y}_{t,f}^{H}}{\Lambda_{t,f}},$$

$$\mathbf{G}_f = \mathbf{R}_f^{-1}\mathbf{r}_f.$$

$$\Lambda_m = \mathrm{NN}(|\mathbf{Y}_m|),$$

$$\Lambda = \textstyle\sum_m \Lambda_m / M$$

- Angle Feature and Fixed Beamforming

$$\mathbf{b}_{i,f} = \mathbf{w}_{i,f}^{H} \mathbf{Y}'_{\mathrm{drv},f}, \qquad \mathbf{a}_{\theta,f} = \sum_{m,n\in\psi} \cos(\mathbf{o}_{mn,f} - \mathbf{r}_{\theta,mn,f})/P,$$

[2] Yihui Fu, Jian Wu, Yanxin Hu, Mengtao Xing, Lei Xie, DESNet: A Multi-channel Network for Simultaneous Speech Dereverberation, Enhancement and Separation, IEEE SLT2021, https://arxiv.org/abs/2011.02131

[5] Keisuke Kinoshita, Marc Delcroix, Haeyong Kwon, Takuma Mori, and Tomohiro Nakatani, "Neural network-based spectrum estimation for online wpe dere- verberation.," in Interspeech, 2017, pp. 384–388.

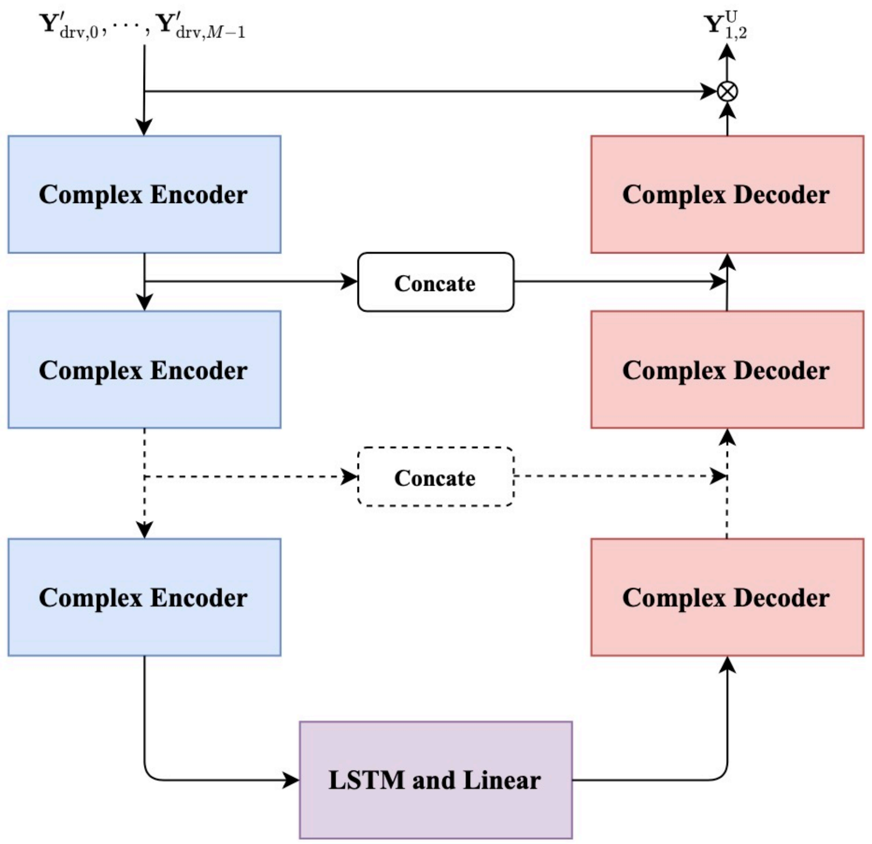
DESNet: A Multi-Channel Network for Simultaneous Speech Dereverberation, Enhancement and Separation

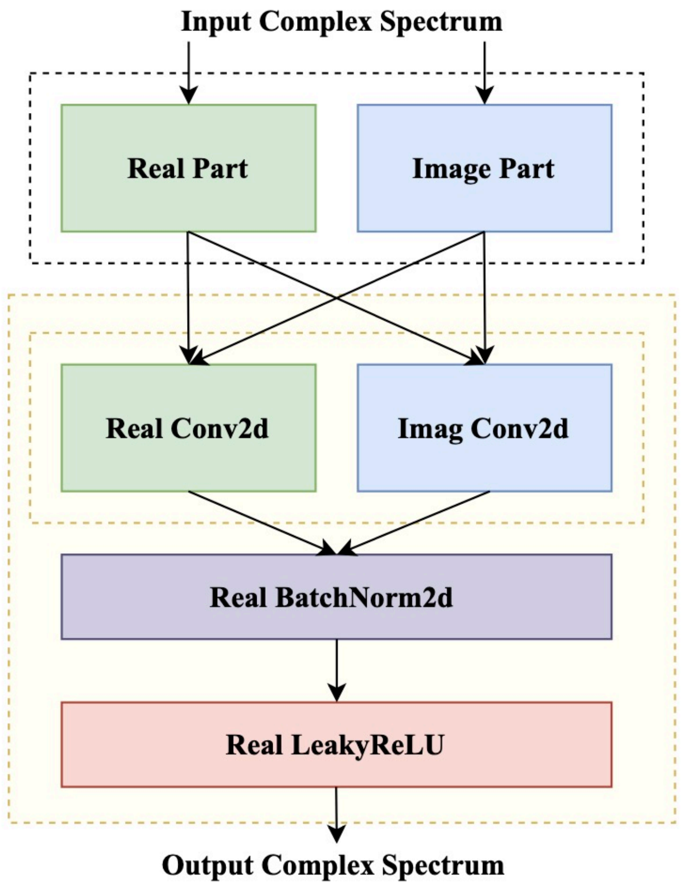
- Speech Unmixing by **DCCRN** [3]
  - A better network can benefit the following selection of the angle and beam features, as well as assist the speech extraction for a better estimation of the final masks
  - **DCCRN** follows the UNet structure, but using **complex-valued convolutional encoders/decoders** and **real/imaginary LSTMs** to model the context dependency.

$$\mathbf{W} \circledast \mathbf{Y} = \begin{bmatrix} \mathbf{W}_r \\ \mathbf{W}_i \end{bmatrix} \circledast \begin{bmatrix} \mathbf{Y}_r \\ \mathbf{Y}_i \end{bmatrix} = \begin{bmatrix} \mathbf{W}_r * \mathbf{Y}_r - \mathbf{W}_i * \mathbf{Y}_i \\ \mathbf{W}_r * \mathbf{Y}_i + \mathbf{W}_i * \mathbf{Y}_r \end{bmatrix},$$

$$\mathbf{M}_{c,\text{mag}} = \tanh\left(\sqrt{\mathbf{H}_{c,r}^2 + \mathbf{H}_{c,i}^2}\right),$$
$$\mathbf{M}_{c,\text{pha}} = \arctan2(\mathbf{H}_{c,i}, \mathbf{H}_{c,r}).$$

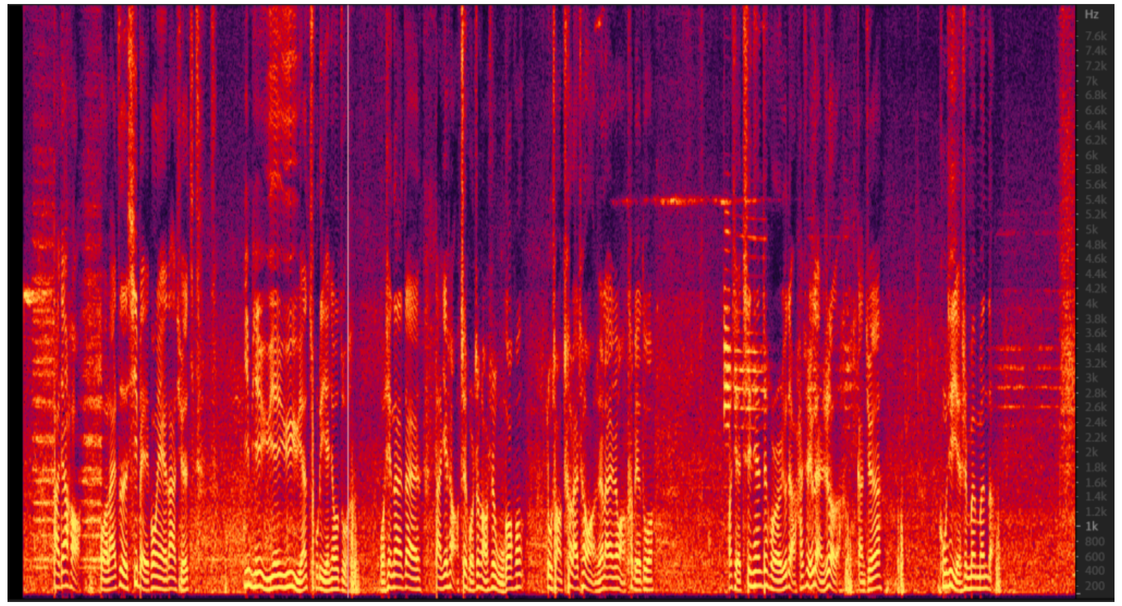
$$\mathbf{Y}_c^U = \mathbf{M}_c^U \odot \mathbf{Y}'_{\text{drv},0}.$$



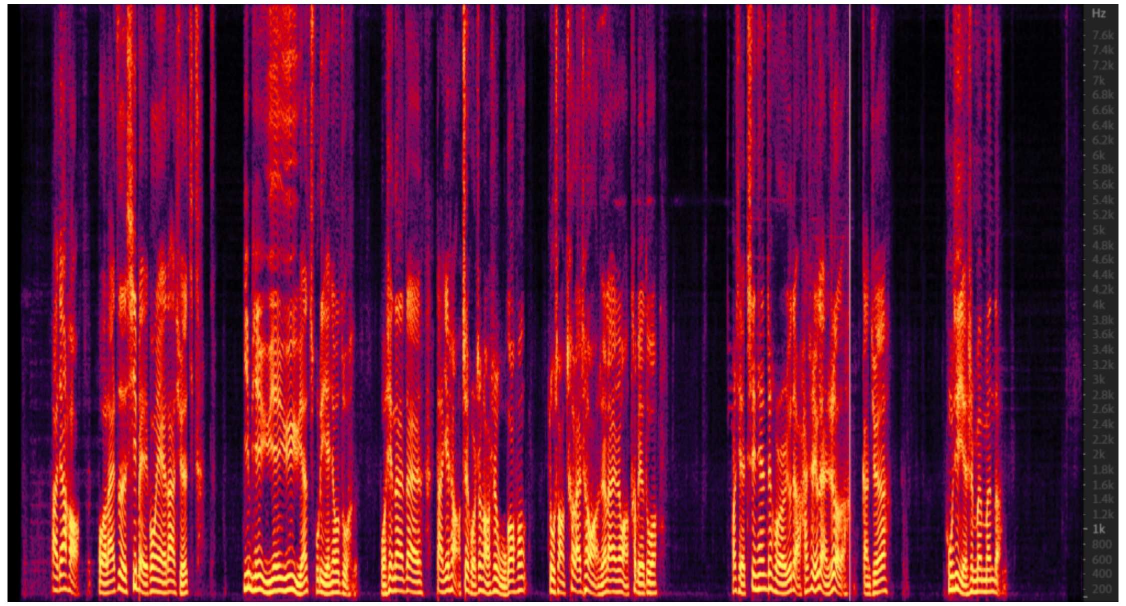


[4] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, Lei Xie, DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement, Interspeech2020, October 25-29, Shanghai, China https://arxiv.org/abs/2008.00264

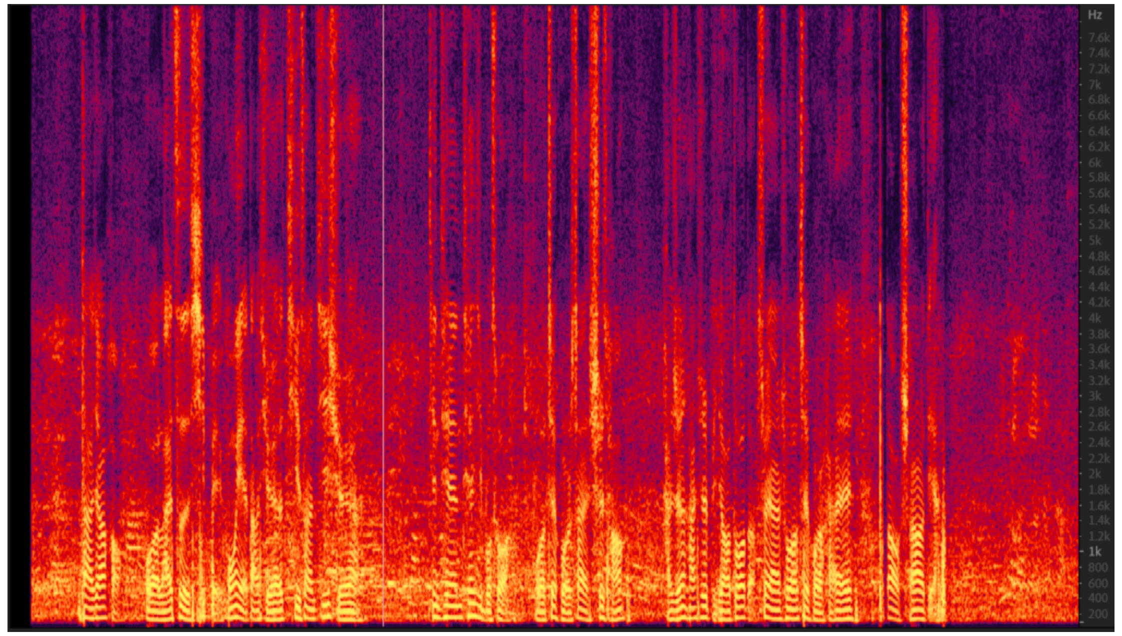
Demo: Speech Enhancement using DCCRN



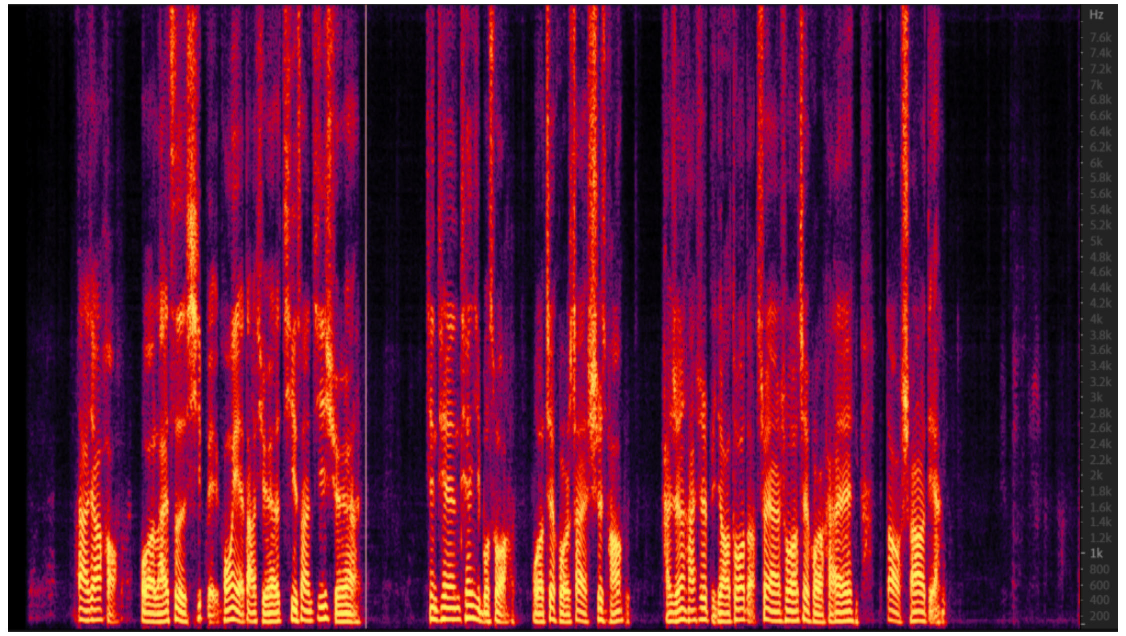
Street-ori



Street-enh



Canteen-ori



Canteen-enh

More samples : https://huyanxin.github.io/DeepComplexCRN/

DESNet: A Multi-Channel Network for Simultaneous Speech Dereverberation, Enhancement and Separation

- Attentional Feature Selection

$$\mathbf{V}_c^U = |\mathbf{M}_c^U|\mathbf{W}_p,$$
$$\mathbf{V}_\theta^A = \mathbf{A}_\theta \mathbf{W}_a,$$

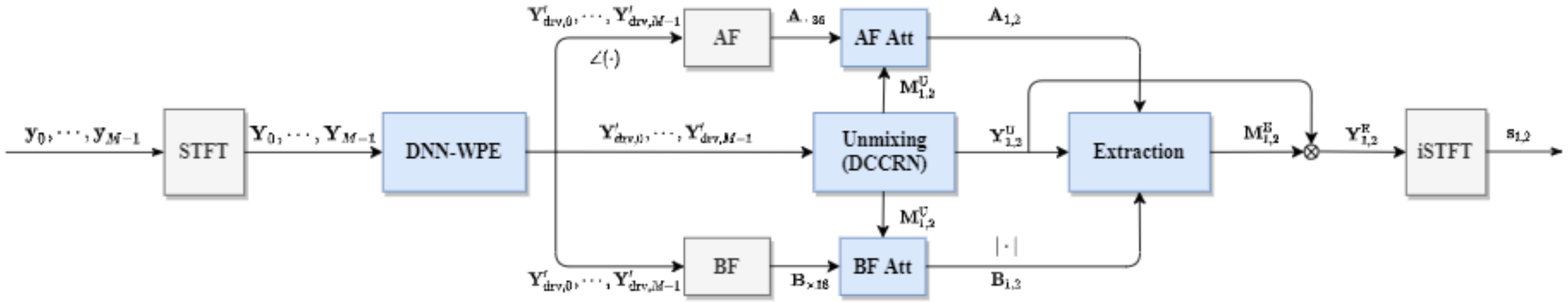
$$s_{c,\theta,t} = (\sqrt{D})^{-1} \left(\mathbf{V}_{c,t}^U\right)^T \mathbf{V}_{\theta,t}^A.$$

$$\hat{s}_{c,\theta} = (T)^{-1} \sum_t s_{c,\theta,t},$$

$$w_{c,\theta} = \mathrm{softmax}_\theta(\hat{s}_{c,\theta}).$$

$$\hat{\mathbf{A}}_c = \sum_\theta w_{c,\theta} \mathbf{A}_\theta.$$

- Speech Extraction
  - Concatenate unmixed speech and attentional features along frequency dimension to estimate the final enhanced and separated mask of two speakers



DESNet: A Multi-Channel Network for Simultaneous Speech Dereverberation, Enhancement and Separation

- Loss Function

$$\text{SI-SNR}(\mathbf{s}_i, \mathbf{x}_j) = 20 \log_{10} \frac{\|\alpha \cdot \mathbf{x}_j\|}{\|\mathbf{s}_i - \alpha \cdot \mathbf{x}_j\|}$$

- Proposed **Symphonic Loss**: the loss calculation of each training chunk in one mini-batch is different
  - If current mixture chunk contains one speaker, namely in SE track, we only optimize the first branch of the network
  - For NSS and CSS tracks, we optimize both branches of the network using permutation invariant training (PIT):

$$\mathcal{L} = -\max_{\phi \in \mathcal{P}} \sum_{(i,j)\in\phi} \text{Si-SNR}(\mathbf{s}_i, \mathbf{x}_j)/N_{\mathcal{P}}$$

- **Staged SNR Strategy**

Table 1. SNR (SDR) range (dB) in each stage.

Training Epoch	SE	CSS	NSS SE	NSS SS
$1 \sim 5$	[5, 10]	[-2, 2]	$\times$	$\times$
$6 \sim 10$	[0, 10]	$\times$	[15, 20]	[-2, 2]
$11 \sim 15$	[-2, 10]	$\times$	[10, 20]	[-4, 4]
$16 \sim 20$	[-5, 10]	$\times$	[5, 20]	[-5, 5]

DESNet: A Multi-Channel Network for Simultaneous Speech Dereverberation, Enhancement and Separation

- Experiments
  - Training & evaluation data
    - On-the-fly data simulation using Librispeech + DNS noise
    - Additional isotropic noise is used
    - Sound source angle : at least 20°
    - Source-Mic distance: 1-5m
    - RT60: 0.1-0.5s
    - Topological structure : 4 mics with 5cm radius
  - Scenario
    - Speech enhancement (SE)
    - Clean source separation (CSS)
    - Noisy source separation (NSS)
  - Feature
    - STFT: 32/16ms
    - Beam number: 18
    - Angle feature number: 36
  - Network Configurations
    - Attention embedding size: 257
    - DCCRN: 6 layers complex CNN
    - Extraction: 3 layers LSTM with 512 hidden size
  - Evaluation metric
    - PESQ for SE
    - Si-SNR for CSS and NSS

DESNet: A Multi-Channel Network for Simultaneous Speech Dereverberation, Enhancement and Separation

- Results

Table 2. Results of non-dereverberated SE and SS.

Model SNR (SDR)	SE (PESQ) -5	0	5	10	Avg.	CSS (SI-SNR (dB)) -5	-2	0	Avg.	NSS (SI-SNR (dB)) -5	-2	0	Avg.
Mixed	1.51	1.87	2.22	2.57	2.04	0.00	0.00	0.00	0.00	-1.63	-0.88	-0.76	-1.09
CACGMM	2.14	2.40	2.69	2.88	2.53	4.50	6.16	6.48	5.71	1.72	4.08	4.46	3.42
Proposed DESNet	**2.55**	**2.87**	**3.17**	**3.41**	**3.00**	**10.18**	**9.98**	**9.78**	**9.98**	**7.16**	**7.73**	**7.77**	**7.55**
- Staged SNR	2.51	**2.87**	3.16	3.40	2.99	9.88	8.54	7.87	8.76	**7.16**	6.65	6.19	6.67
- Symphonic Loss	2.36	2.73	3.06	3.33	2.87	9.61	9.40	9.26	9.42	6.70	7.31	7.31	7.11
- BF Feature	2.29	2.65	2.97	3.23	2.79	8.77	8.65	8.44	8.62	5.84	6.32	6.31	6.16
DCCRN	2.25	2.61	2.94	3.20	2.75	7.78	6.04	5.37	6.40	5.73	4.62	4.07	4.81
Conv-TasNet	2.00	2.29	2.53	2.71	2.38	6.03	6.67	6.72	6.47	3.93	5.09	5.23	4.75
DPRNN	2.22	2.55	2.84	3.09	2.68	9.09	9.36	9.32	9.26	6.37	7.32	7.42	7.04
FasNet	2.24	2.58	2.89	3.14	2.71	9.42	9.35	9.02	9.26	6.91	7.63	7.41	7.32

Table 3. Results of dereverberated SE and SS.

Model SNR (SDR)	SE (PESQ) -5	0	5	10	Avg.	CSS (SI-SNR (dB)) -5	-2	0	Avg.	NSS (SI-SNR (dB)) -5	-2	0	Avg.
Mixed	1.41	1.71	2.02	2.31	1.86	-1.38	-0.75	-0.64	-0.92	-2.63	-1.54	-1.35	-1.84
CACGMM	2.09	2.36	2.63	2.83	2.48	3.97	5.54	5.85	5.12	1.57	3.90	4.27	3.25
Proposed DESNet	**2.36**	**2.65**	**2.90**	**3.12**	**2.76**	**8.07**	**8.18**	**8.14**	**8.13**	**6.38**	**6.65**	**6.50**	**6.51**
- Staged SNR	2.26	2.57	2.84	3.06	2.68	7.96	8.14	8.03	8.04	5.56	6.36	6.18	6.03
- Symphonic Loss	2.32	2.63	2.89	3.11	2.74	7.74	7.88	7.42	7.68	5.68	6.45	6.50	6.21
- DNN-WPE	2.17	2.49	2.77	3.01	2.61	7.36	7.66	7.59	7.54	5.20	5.68	5.65	5.51
WPE-DCCRN	2.16	2.49	2.78	3.00	2.61	6.64	6.09	5.77	6.17	5.16	5.07	4.61	4.95

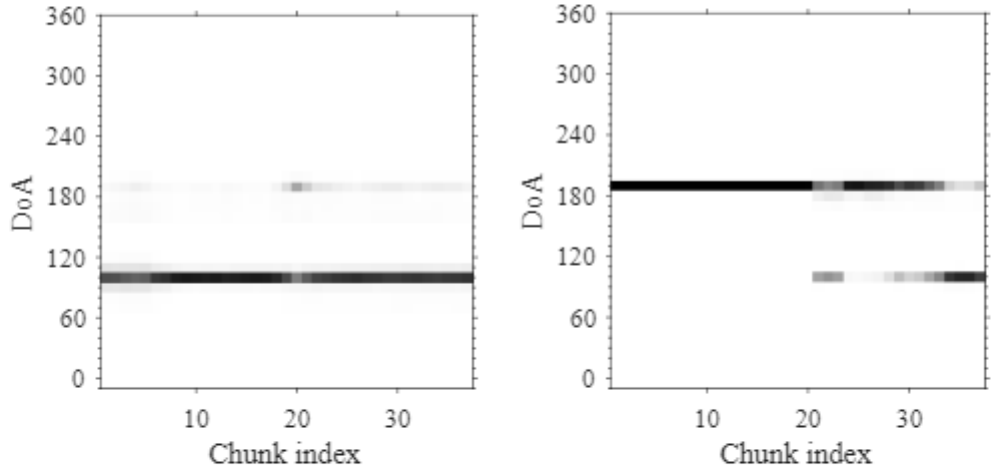


Fig. 4. Example of the learnt weights on angle feature in a two-speaker mixture utterance.

- In both non-dereverberated and dereberberated SE and SS scenarios, DESNet suppress CACGMM, DCCRN and time domain approaches including Conv-Tasnet, DPRNN and FasNet
- Staged SNR, symphonic loss and BF Feature are effective for better enhancement and separation performance
- The learnt attentional weight fits the actual speaker's direction perfectly
- Future work: optimizing speech dereverberation, enhancement and separation with acoustic model to further improve the speech recognition accuracy in real environment scenarios

Demo: DesNet

		SE	CSS		NSS	
Non-dereverberated	Input					
	Output					
Dereverberated	Input					
	Output					

[2] Yihui Fu, Jian Wu, Yanxin Hu, Mengtao Xing, Lei Xie, DESNet: A Multi-channel Network for Simultaneous Speech Dereverberation, Enhancement and Separation, IEEE SLT2021, https://arxiv.org/abs/2011.02131

More demos: https://felixfuyihui.github.io/DesNet_Demo/

Thanks!

西工大音频语音与语言处理研究组

Follow us thru Wechat

Visit us at www.npu-aslp.org