Tencent AI Lab

Keynote at CHIME 2020

Solving Cocktail Party Problem
- From Single Modality to Multi-Modality

Dong Yu
Tencent AI Lab

Enable High-Accuracy ASR in Everyday Environments

Tencent AI Lab

Outline

- ❑Mask-based -> Filter-based
- ❑Blind Source Separation -> Target Speaker Extraction
- ❑Single Modality -> Multi-Modality
- ❑Summary

Outline

- **Mask-based -> Filter-based**
- Blind Source Separation -> Target Speaker Extraction
- Single Modality -> Multi-Modality
- Summary

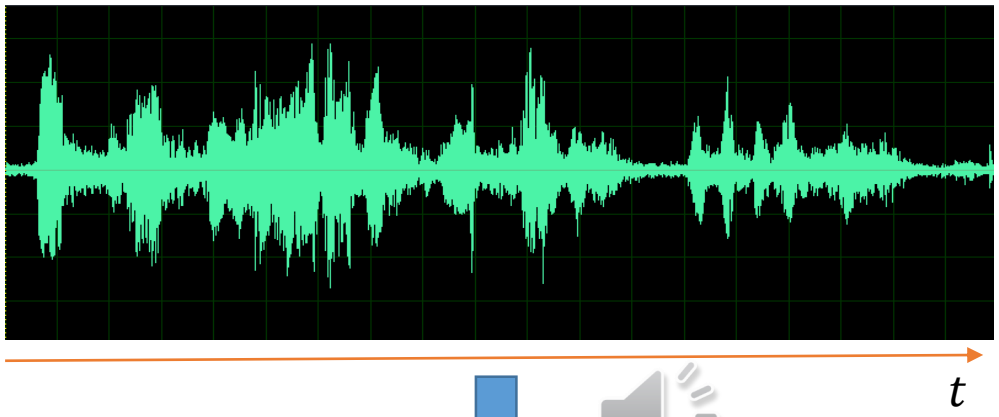
Speech Separation / Enhancement / Dereverberation

Tencent AI Lab

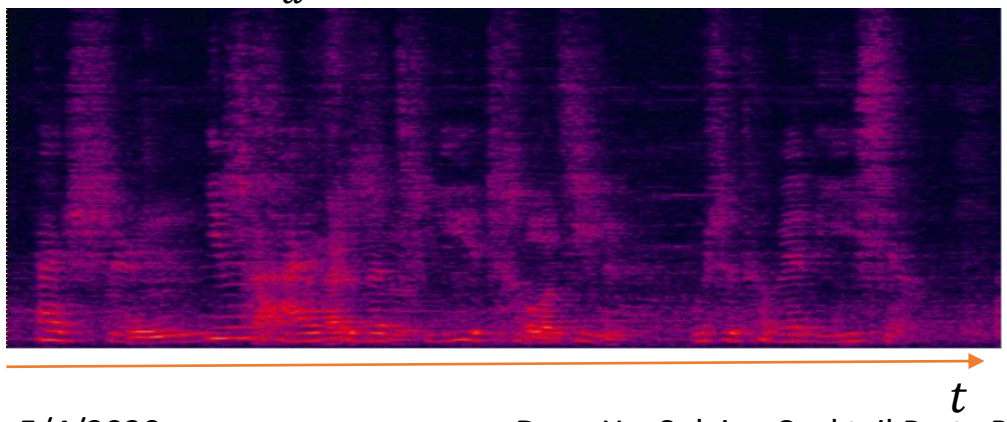
Mixture Signal at M Mics

Target speech + interfering speech + noise + reverberation

$$y^{(m)} = s_d * h^{(m)} + n^{(m)} = s^{(m)} + n^{(m)}$$



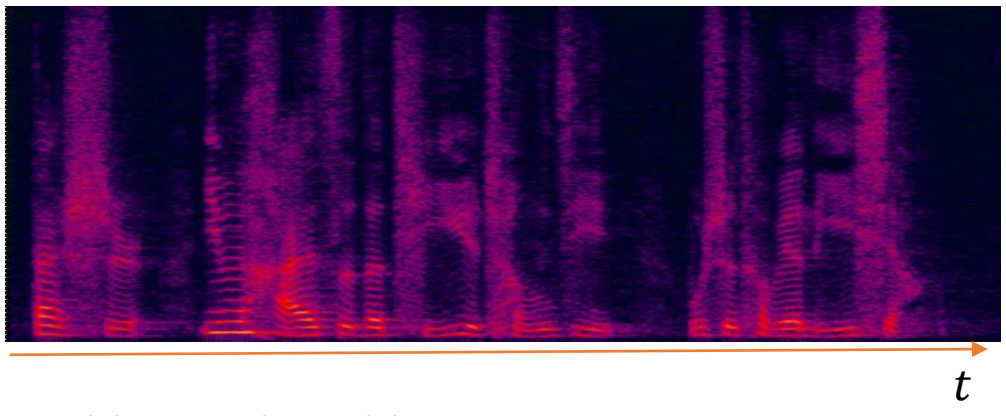
$$Y^{(m)} = S_d \cdot H^{(m)} + N^{(m)} = S^{(m)} + N^{(m)}$$



Dry Clean Target Speech s_d

(Or Reverb Clean Target Speech $s_d * h^{(m)}$)

S_d or $S_d \cdot H^{(m)}$



Ideal Solution

Tencent AI Lab

Realtime and High Performing

for both communication and back-end applications

Popular Solution: Estimate and Apply a Mask/Gain

Tencent AI Lab

Scalar $= f_{t,f}(Y)$: One Mask per time-frequency bin

Reverb Clean Speech (remove additive noise)

Mixture Signal

$$S_{t,f} = \boldsymbol{Mask}_{t,f} \times Y_{t,f}$$

(1) Estimating masks **generalize better** than estimating spectrum directly

(2) Masks may be estimated with deep learning or signal processing techniques

(3) Masks are often estimated by **exploiting spatial, frequency, and history information**

Popular Solution: Estimate and Apply a Mask/Gain

Tencent AI Lab

❑ Explicitly estimate the masks

❖ Ideal Ratio Mask

$$\text{IRM(t,f)} = \frac{|S^2(t,f)|}{|S^2(t,f)| + |N^2(t,f)|}$$

IRM$\in [0, 1]$

(1) Not well defined at silence segments

❖ Minimize

$$\mathcal{L}1 = \text{MSE}(\widehat{\text{IRM}}, \text{IRM})$$

(2) Not optimizing for the final criterion

❖ Recover reverb clean

$$|S^2(t,f)| = \widehat{\text{IRM}} \cdot (|S^2(t,f)| + |N^2(t,f)|) \approx \widehat{\text{IRM}} \cdot |Y^2(t,f)|$$

(3) Requires Independence assumption **(Poor esp. when SNR is low)**

Popular Solution: Estimate and Apply a Mask/Gain

Tencent AI Lab

❑**Solution**: treat masks as hidden variables

❖ $\mathcal{L}2 = \mathbf{SISNR}\left(\text{wav}_{\text{ref}}, \text{wav}_{pred}(\sigma(Y), Y)\right)$ (for **IRM**$\in [\mathbf{0}, \mathbf{1}]$)

(2) Optimize SISNR defined on estimated waveforms

(1) Sigmoid masks may not exactly align with IRMs

❑Better Solution: Define ReLU hidden amplitude masks

❖ $\mathcal{L}3 = \mathbf{SISNR}(\text{wav}_{\text{ref}}, \text{wav}_{\text{pred}}(\text{ReLU}(Y), Y))$

for $\text{ReLU-Mask(t,f)} = \dfrac{|S(t,f)|}{|Y(t,f)|}$

(1) Phase is missing

$\mathbf{ReLU - Mask}(\boldsymbol{t}, \boldsymbol{f}) \in [\mathbf{0}, +\infty]$

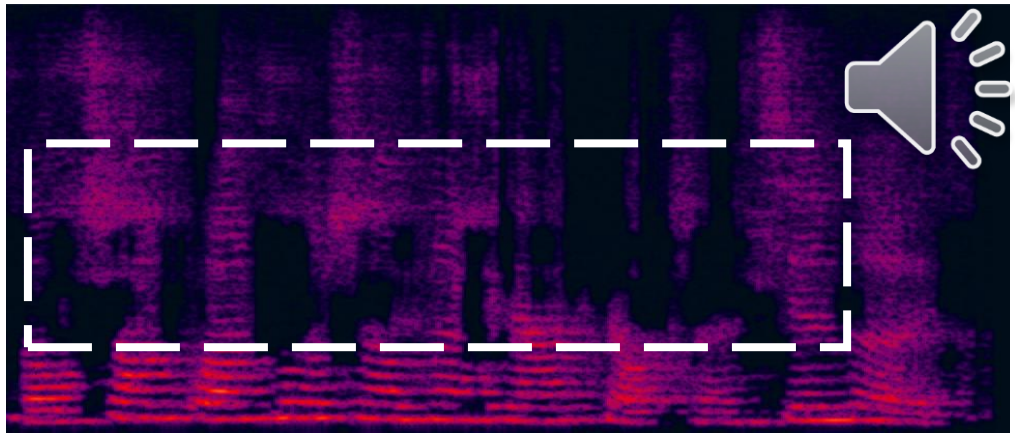
(2) Tend to get masks with values close to 0 -> black hole

Use Complex-Valued Masks (the Choice)

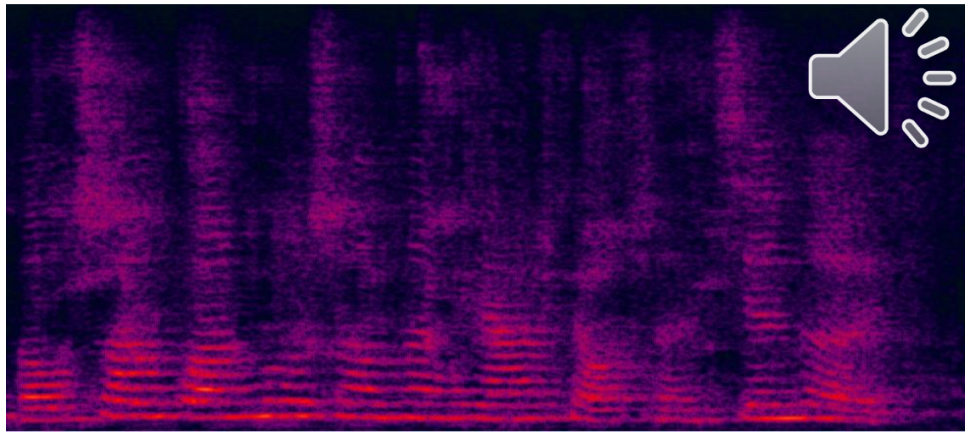
$$S_r + iS_i = (M_r + iM_i) * (Y_r + iY_i)$$
$$= (M_r Y_r - M_i Y_i) + i(M_r Y_i + M_i Y_r)$$

$M_r, M_i \in [-\infty, +\infty]$
No need to clip the range as in earlier works since masks here are hidden variables

- Estimate both magnitude and phase
- Can avoid "spectral black holes" distortion observed in sigmoid/relu masks and achieve better hearing perception and better ASR performance



ReLU mask (WER 17.7%)



Complex mask (WER 16.9%)

Extend to Multi-Channel (Spatial Filtering)

Tencent AI Lab

- Minimum Variance Distortionless Response (MVDR) Filter

$$S_{t,f} = \mathbf{W}_{t,f}^{H} \mathbf{Y}_{t,f}$$

$\mathbf{W}_{t,f}$ and $\mathbf{Y}_{t,f}$ are M-dim complex vectors with M mics

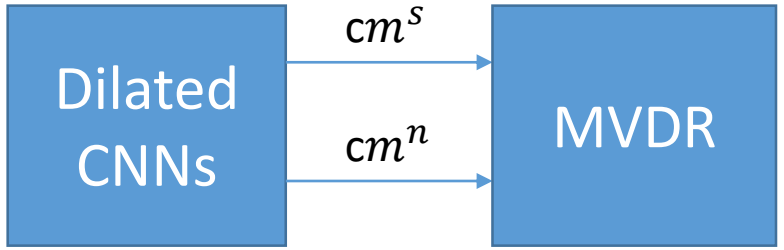
$$\mathbf{W} = \frac{\Phi_{NN}^{-1} v}{v^H \Phi_{NN}^{-1} v} \text{ or } \mathbf{W} = \frac{\Phi_{NN}^{-1} \Phi_{SS}}{trace(\Phi_{NN}^{-1} \Phi_{SS})} u$$

Desired direction

Result of optimizing $\widehat{W} = \underset{W}{\operatorname{argmin}} \sum_t |\mathbf{W}^H \mathbf{Y}_t|^2$ s.t. $\mathbf{W}^H v = 1$

- We can use DL models to estimate hidden complex mask cm^S per (t,f,m)

$$\Phi_{ss} = \frac{1}{\sum_1^T ((cm^S)^H (cm^S))} \sum_1^T (cm^S Y)(cm^S Y)^H)$$



Filtering Across Channels & Frames: Multi-tap MVDR

Tencent AI Lab

- **Core Idea**: Past frames can help estimate current frame

$$S_{t,f} = \bar{\mathbf{W}}_{t,f}^{H} \bar{\mathbf{Y}}_{t,f}$$

(1) Valid for single/multi-channel and multimodal setups

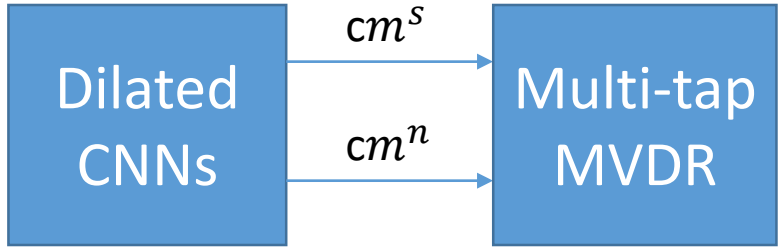
$$\bar{Y}_{t,f} = [Y_{1,t}, \dots, Y_{M,t}, Y_{1,t-1}, \dots, Y_{M,t-1}, \dots\dots, Y_{1,t-L}, \dots, Y_{M,t-L}]$$

(2) Augmented dimensions across channels and frames

$$\hat{\bar{\mathbf{w}}} = \frac{\boldsymbol{\Phi}_{\bar{\mathbf{N}}\bar{\mathbf{N}}}^{-1} \boldsymbol{\Phi}_{\bar{\mathbf{S}}\bar{\mathbf{S}}}}{\mathrm{trace}(\boldsymbol{\Phi}_{\bar{\mathbf{N}}\bar{\mathbf{N}}}^{-1} \boldsymbol{\Phi}_{\bar{\mathbf{S}}\bar{\mathbf{S}}})} \bar{\mathbf{u}}$$

Result of optimizing $\quad \hat{\bar{W}} = \underset{\bar{W}}{\mathrm{argmin}} \sum_{t} \left| \bar{W}_t^H \bar{Y}_t \right|^2 \quad$ s.t. $\quad \mathbf{W_0}^H \mathbf{v}=1$

$$\Phi_{\bar{S}\bar{S}} = \sum_{t=1}^{T} \frac{(\bar{S})(\bar{S})^H}{\sum_1^T ((cm^s)^H (cm^s))} = \sum_{t=1}^{T} \frac{(\overline{cm}^s \bar{Y})(\overline{cm}^s \bar{Y})^H}{\sum_1^T ((cm^s)^H (cm^s))}$$



MVDR with Complex Masks (PESQ on 500-utt Test Set)

Tencent AI Lab

Upper: single channel direct mask estimation.
Lower: multi channel. mask is estimated to compute the covariance matrix in MVDR. More friendly to ASR

Angle between target and interfering speakers

Number of overlapping speakers

Reverb Clean WER=7.0%

	<15°	<45°	<90°	<180°	2 SPK	3 SPK	1 SPK	AVE	WER
Mixture	1.88	1.88	1.98	2.03	2.02	1.77	3.55	2.16	51.3%
Sigmoid Mask	1.78	1.96	2.20	2.08	2.14	1.74	3.62	2.22	27.7%
ReLU Mask	2.54	2.73	2.92	2.88	2.85	2.56	**3.89**	2.91	17.7%
Complex Mask	**2.64**	**2.84**	**3.00**	**3.00**	**2.94**	**2.66**	**3.89**	**3.00**	**16.9%**
Sigmoid Mask w/MVDR	2.24	2.56	2.80	2.72	2.64	2.35	3.67	2.71	16.7%
ReLU Mask w/MVDR	2.52	2.74	2.94	2.85	2.86	2.54	3.68	2.88	12.6%
Complex Mask w/MVDR	**2.55**	**2.77**	**2.97**	**2.89**	**2.89**	**2.57**	**3.73**	**2.91**	**11.8%**
CM Multi-tap MVDR	**2.68**	**2.97**	**3.17**	**3.10**	**3.07**	**2.74**	3.79	**3.07**	**10.2%**

Multi-tap MVDR vs MVDR (PESQ on 5007-utt Test Set)

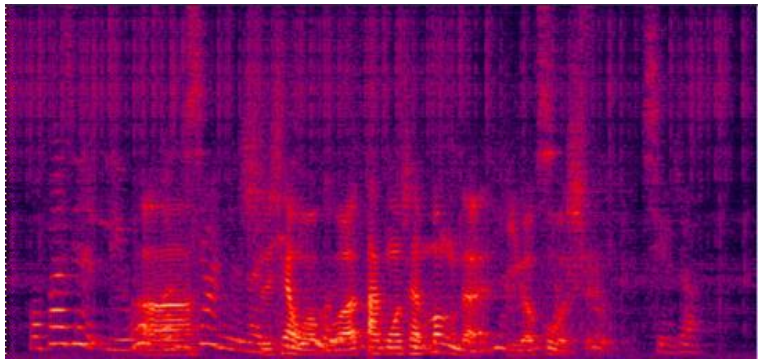
Tencent AI Lab

Angle between target and interfering speakers

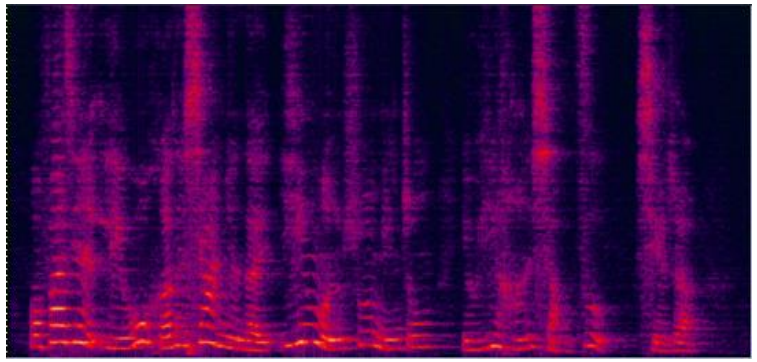
Number of overlapping speakers

MVDR	<15°	<45°	<90°	<180°	2 SPK	3 SPK	1 SPK	AVE
PESQ	2.68	2.87	3.01	3.01	2.94	2.75	3.76	2.89
SDR	8.86	10.85	12.55	12.71	11.54	9.73	20.48	11.08
SISNR	7.45	9.40	11.00	11.22	10.06	8.29	18.08	9.6
WER				16.6%				
Multi-tap MVDR	**<15°**	**<45°**	**<90°**	**<180°**	**2 SPK**	**3 SPK**	**1 SPK**	**AVE**
PESQ	**2.75**	**2.99**	**3.15**	**3.14**	**3.05**	**2.85**	**3.80**	**2.99**
SDR	**9.57**	**12.20**	**14.12**	**14.04**	**12.80**	**10.82**	**20.90**	**12.22**
SISNR	**7.92**	**10.44**	**12.28**	**12.28**	**10.98**	**9.15**	**18.16**	**10.42**
WER				**15.0%**				

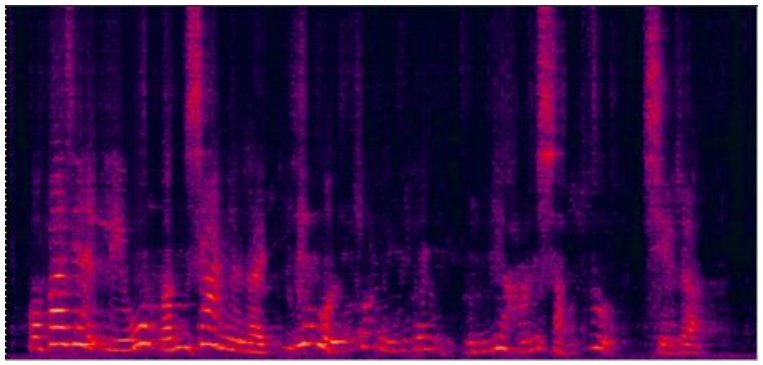
Multi-tap MVDR: Example



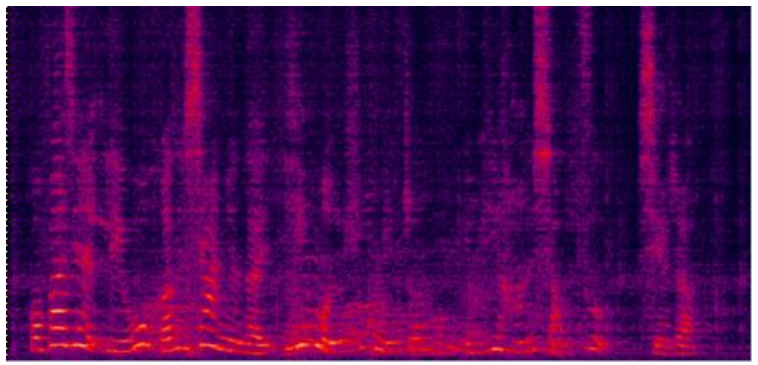
2 speakers mix + noise



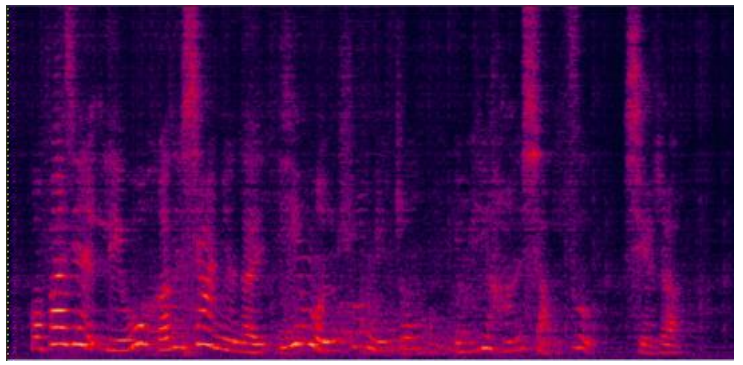
Reverb Clean



Complex Mask
(single channel)



Conventional MVDR
with Complex Mask



Multi-tap MVDR
with Complex Mask

Further Works Not Covered Here

Tencent AI Lab

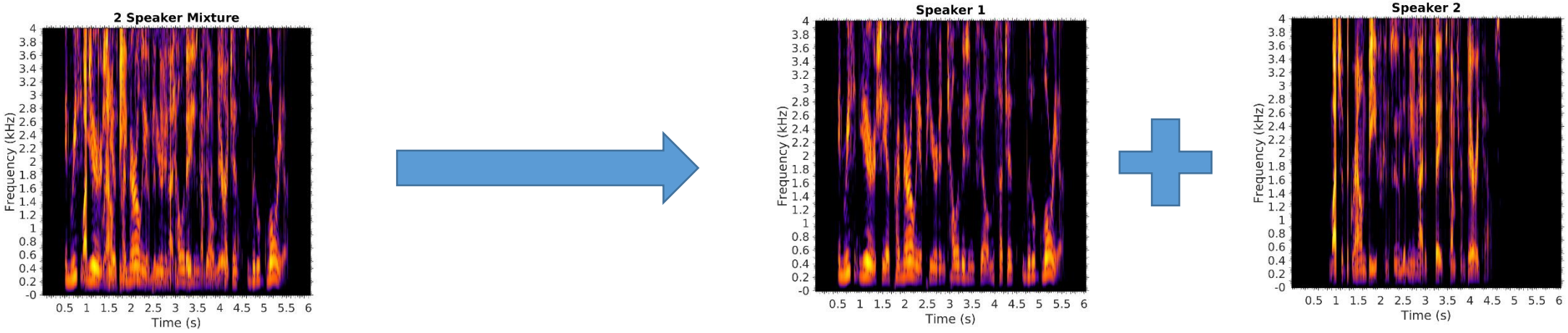
- Filtering across channels, frames & frequency bins
  - Because STFT causes frequency leakage when the window size is not large enough
- Conduct separation/enhancement/dereverberation simultaneously
  - So that we can recover dry clean speech (I only covered reverb clean in this talk)
  - Exploitation of frames that cover both early and late reverberation
  - Extend Weighted Power minimization Distortionless response (WPD)
- Direct estimation of filter weights $\mathbf{W}$ esp. with multi-task joint training
  - I only discussed the approach of estimating the masks as a way to estimate covariance matrix in MVDR

We will disclose related techniques and results in upcoming papers

Outline

- ❑ Mask-based -> Filter-based
- ❑ **Blind Source Separation -> Target Speaker Extraction**
- ❑ Single Modality -> Multi-Modality
- ❑ Summary

Blind Source Separation



- Allows systems to listen to multiple speakers simultaneously
- Learns from training set to look for hidden regularities (complicated soft constraints)
- Problem: **Label Ambiguity** / Label Permutation Problem
- Solutions: **deep clustering** and **permutation invariant training**

Target Speaker Extraction

When there is information to identify the speaker of interest

Important in many application scenarios

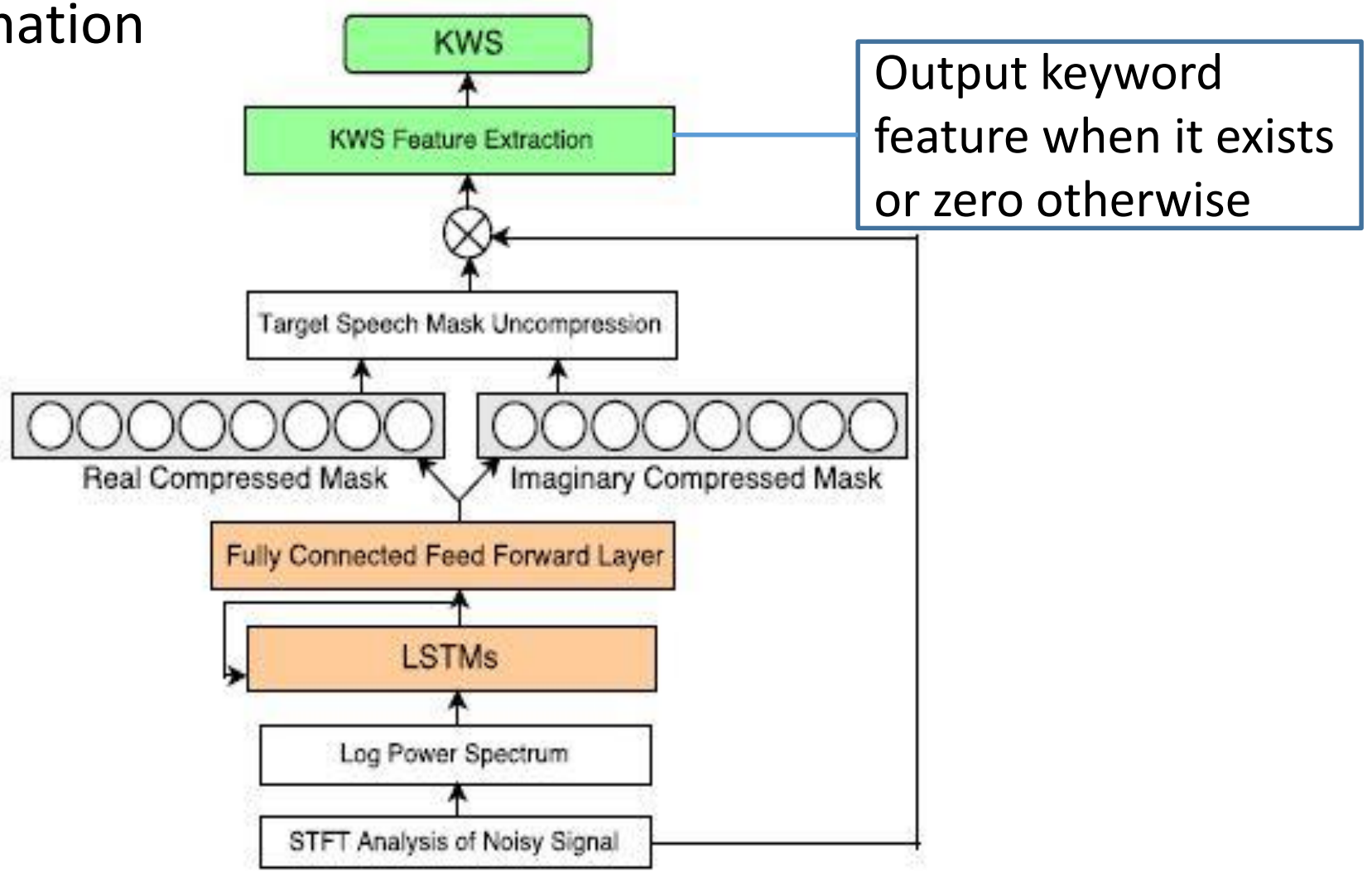
Find information that can identify the target speaker

Keyword Spotting: Speaker Who Says the Keyword

Keyword is the information to identify the target speaker

Single channel

Text-dependent & speaker-independent



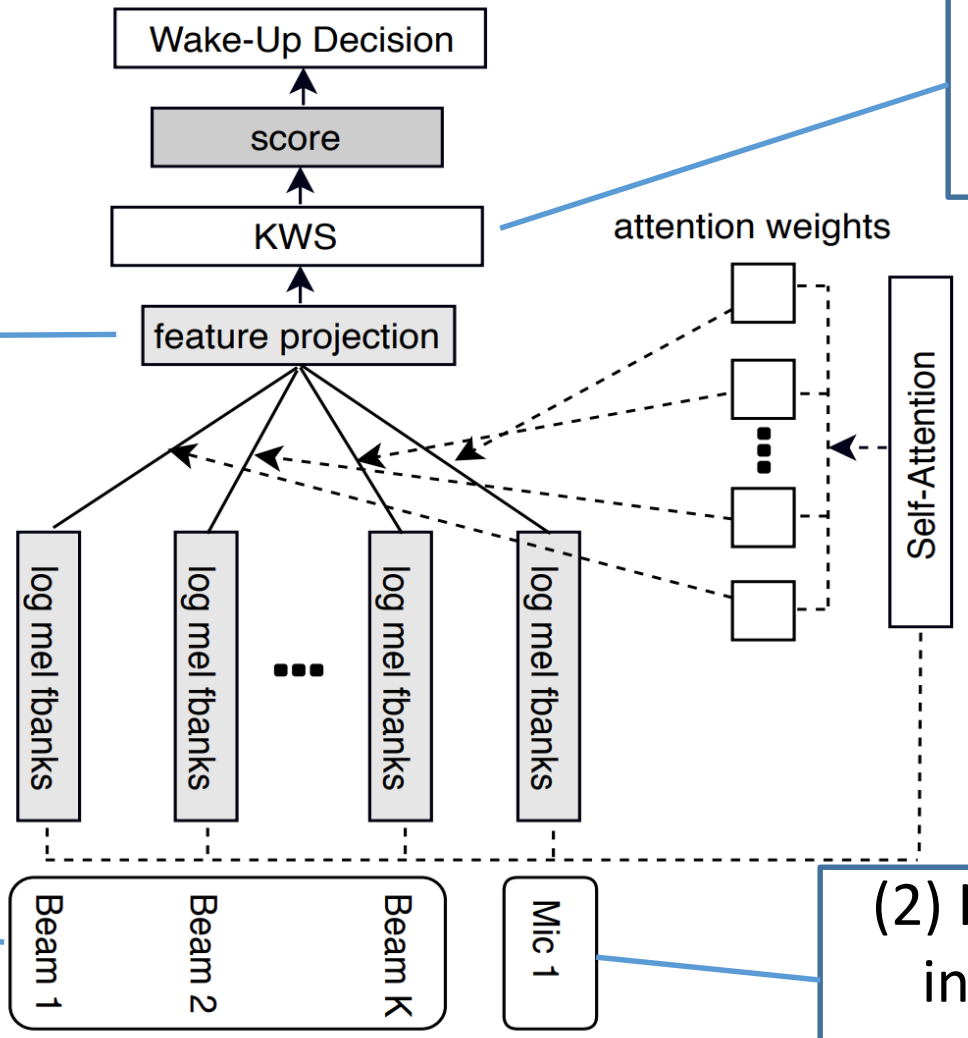
Beamformer based Multi-look Enhancement

Tencent AI Lab

Multi-Channel

DOA independent

(3) weighted sum using attention mechanism



(4) Joint training improves performance

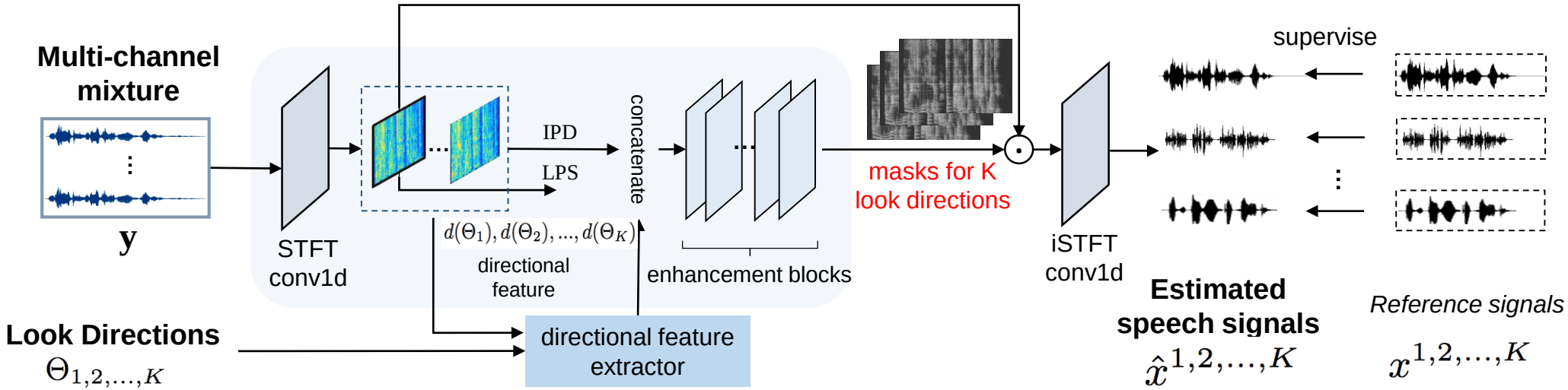
(1) Fixed beamformers with multiple fixed beams, equally sampled in space

(2) Keep one raw mic input to preserve target speech quality

Multi-look Neural Enhancement

Further Improvement

over beamformer based multi-look enhancement



Each neural enhancer enhances the speech source closest to each look direction

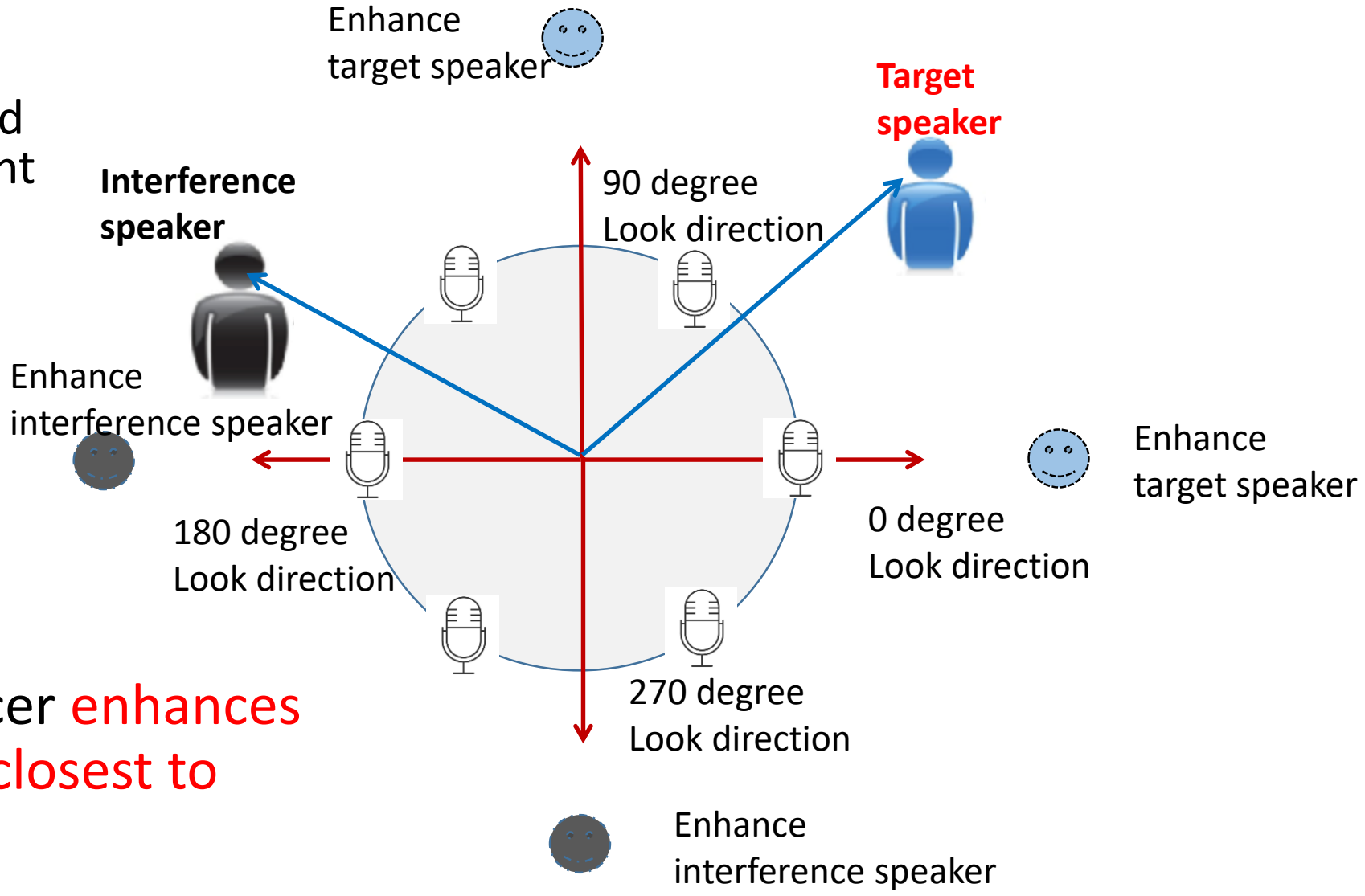
$$x^j = x_{\tilde{j}} \quad \tilde{j} = \arg\min_i |\Theta_j - \theta_i|$$

$$j = 1, 2, ..., K, \; i = 1, 2, ..., N$$

Multi-look Neural Enhancement

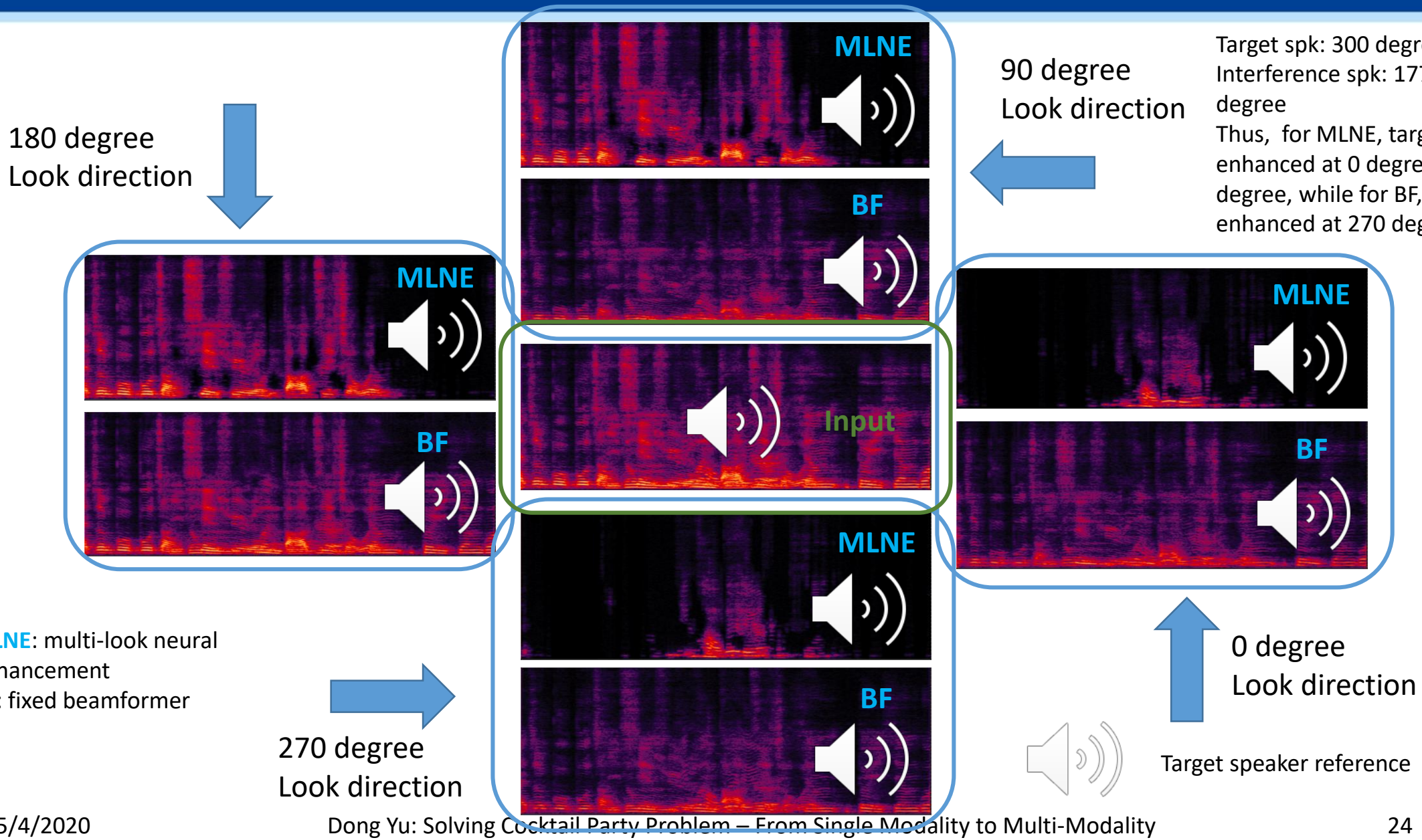
Further Improvement

over beamformer based multi-look enhancement



Each neural enhancer enhances the speech source closest to each look direction

Comparison with Beamformer based Multi-Look



Target spk: 300 degree
Interference spk: 177 degree & 209 degree
Thus, for MLNE, target speaker enhanced at 0 degree and 270 degree, while for BF, target speaker enhanced at 270 degree

MLNE: multi-look neural enhancement
BF: fixed beamformer

Keyword Spotting Accuracy Comparison

Tencent AI Lab

	-6 ~ 6dB SIR (multi-talker)	>6dB SIR (multi-talker)	One speaker SNR>6db
Raw input	10.98%	61.41%	83.16%
Beamformer ML Individual Enhancer	62.93%	85.33%	92.60%
Beamformer based multi-look (4 look directions + 1mic)	38.88%	92.92%	**95.09%**
Multi-Look neural enhancement (4 look directions)	88.68%	88.98%	88.95%
Multi-Look neural enhancement (4 look directions + 1 mic)	**93.40%**	**94.45%**	94.02%

Operation point

<=1 false alarm over 12 hours' TV, conversational speech and other noises

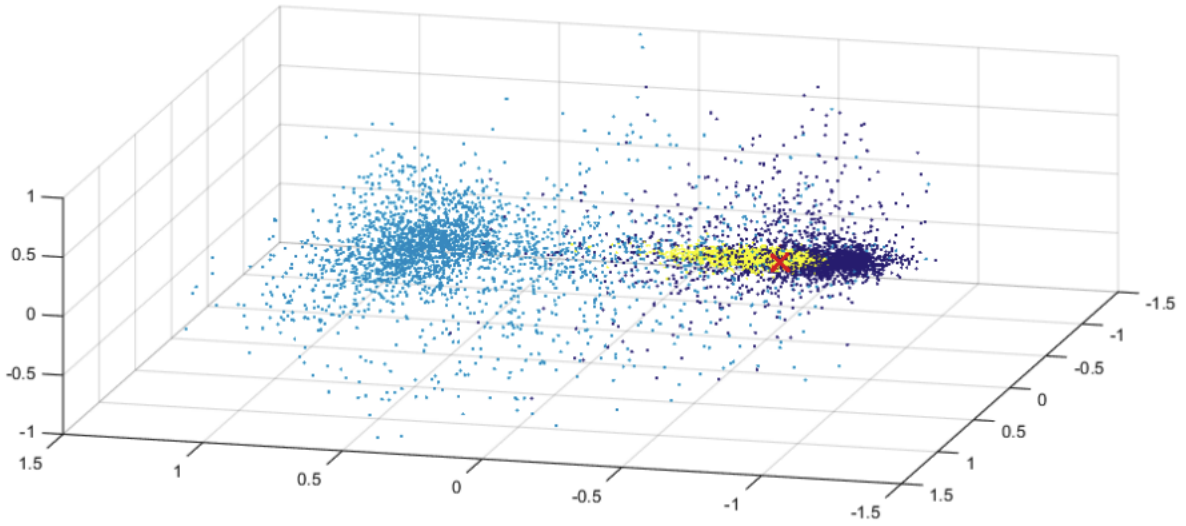
Exploit the Speaker's Voice Characteristics

Tencent AI Lab

- Speaker-aware enhancement
  - When the speaker's voice characteristics is available
  - Address the problem of who to listen to
- Things to leverage
  - users' wakeup words before voice command
  - registered user voice profile on devices

Deep Extractor Network (DENet)

- Extract the speaker whose representation in a canonical space is closer to that of the anchor (e.g., keyword)
- A variant of deep attractor network (DANet)
- **Difference**: Exploit the anchor + the relationship of anchor and the noisy target speech in the canonical space



Location of extractors and embeddings of T-F bins in the new embedding space. Each dot visualizes the first three principle components; Yellow dots indicate canonical extractors of all training speakers, and their centroid is marked with a red cross; Light blue dots indicate embeddings for an interfering speaker, and dark blue dots for a target speaker.

Deep Extractor Network (DENet)

Tencent AI Lab

2-speaker mixture

Model	SDR	PESQ
Orig Mixture	0.99	1.96
DANet Nearest	15.71	2.51
DANet Oracle	16.67	2.58
DANet Anchor	17.14	2.72
DENet	**17.53**	**2.75**

3-speaker mixture

Model	SDR	PESQ	SDR	PESQ
Orig Mix.	0.44	1.87	−2.05	1.46
DANet Near.	11.99	2.30	8.84 (26%)	2.13
DANet Orac.	11.98	2.27	8.85 (26%)	2.10
DANet Anch.	13.32	2.48	10.39 (22%)	2.12
DENet	**13.44**	**2.52**	**10.67 (20%)**	**2.14**

DENet: the proposed method.

DANet: Deep attract network. ”Anchor”: attractor calculated from anchor speech. ”Near”: attractor from the mixture itself which is closer to the attractor from anchor. “Oracle”: original DANet w/o using anchor speech, the largest SDR/PESQ among all speakers in the mixture.

Input



Output



Speaker Embedding as Auxiliary Input: Voice Filter

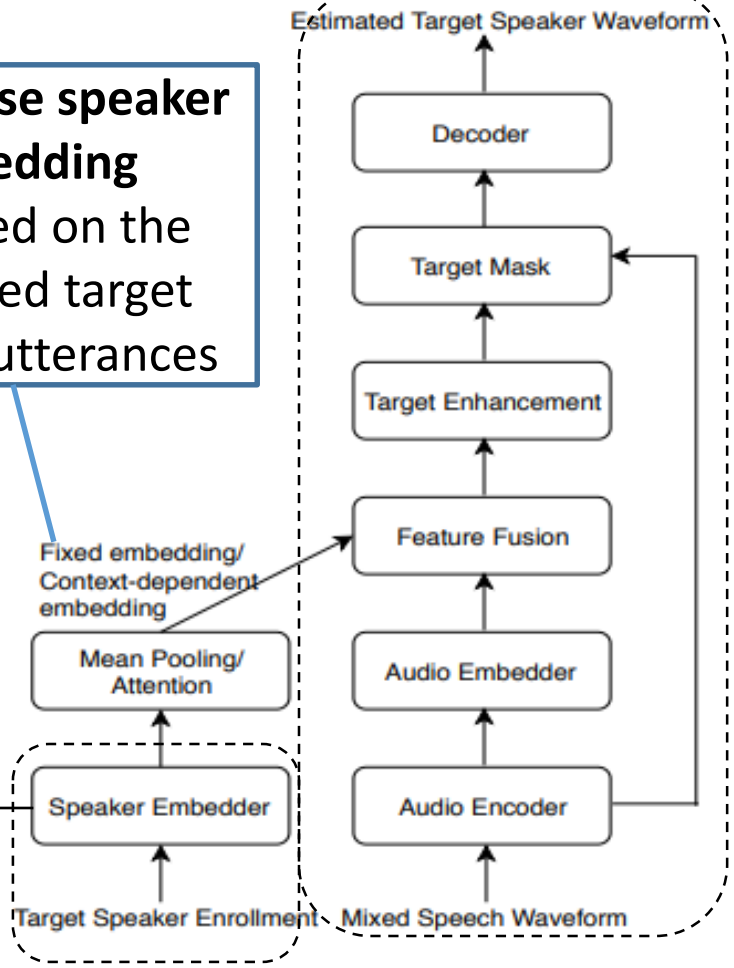
Tencent AI Lab

Original Voice Filter

Enhancement
Si-SNR loss

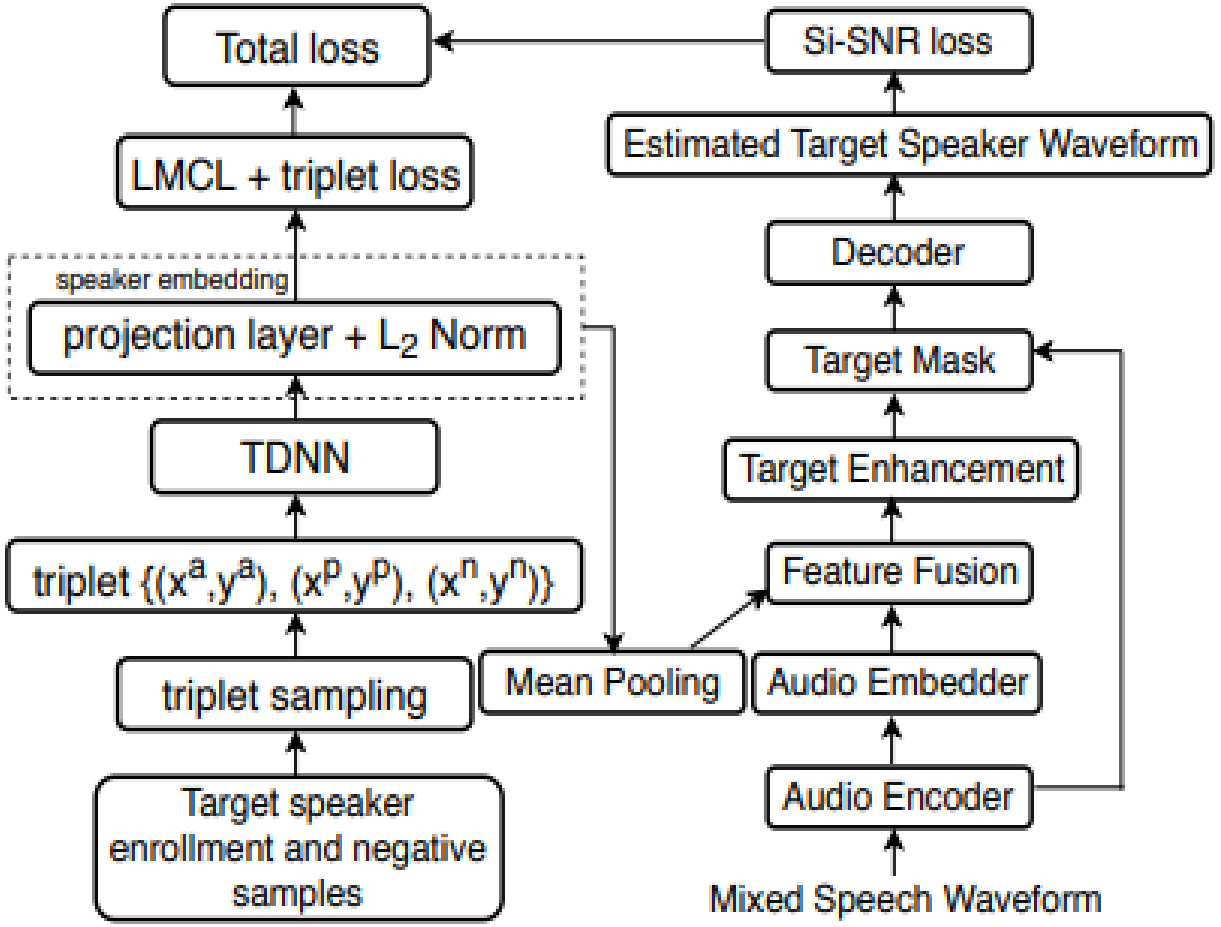
frame-wise speaker embedding extracted on the registered target speaker utterances

Speaker Verification
LMCL+triplet loss



Joint optimization of total loss

Improved Voice Filter



Speaker Embedding as Auxiliary Input: Voice Filter

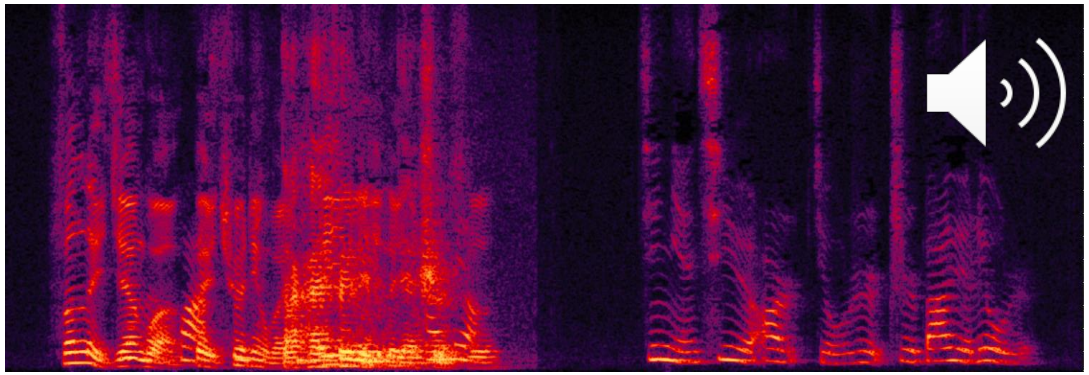
The comparison of different methods in SDR (dB)

Method	1 spk	2 spks	3 spks	Avg.
Raw	13.17	-0.73	-1.74	3.69
Pre-train: mean pooling	21.11	8.37	6.55	12.76
Pre-train: attention	21.31	9.42	7.88	13.62
Finetune: mean pooling	21.19	9.77	7.97	13.71
Joint-train: mean pooling	**21.46**	**10.56**	**9.24**	**14.47**

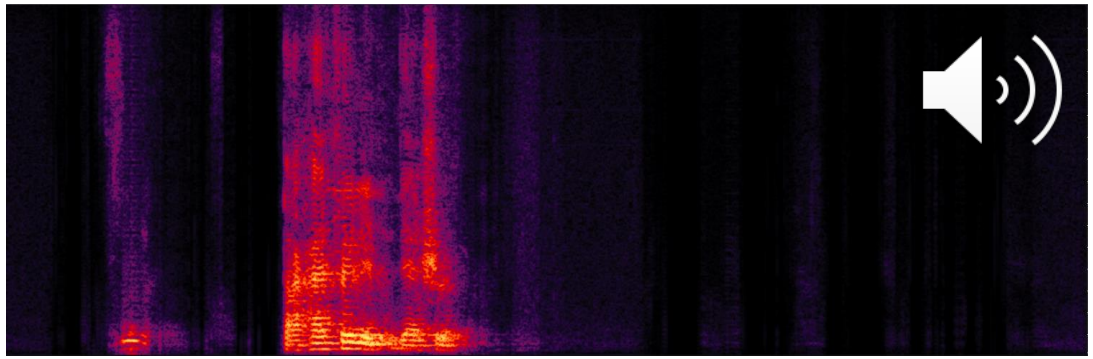
The comparison of different methods in SI-SNR (dB)

Method	1 spk	2 spks	3 spks	Avg.
Raw	13.50	-0.80	-1.87	3.74
Pre-train: mean pooling	20.80	7.82	6.07	12.32
Pre-train: attention	21.00	8.88	7.38	13.17
Finetune: mean pooling	20.89	9.17	7.44	13.25
Joint-train: mean pooling	**21.17**	**9.98**	**8.66**	**14.00**

Joint-train: the proposed method. Pre-train: pre-trained speaker verification model used for Voice Filter.
Finetune: pretrained speaker verification model finetuned on AISHELL-2 data. (AISHELL-2 is used for joint-training model)



input

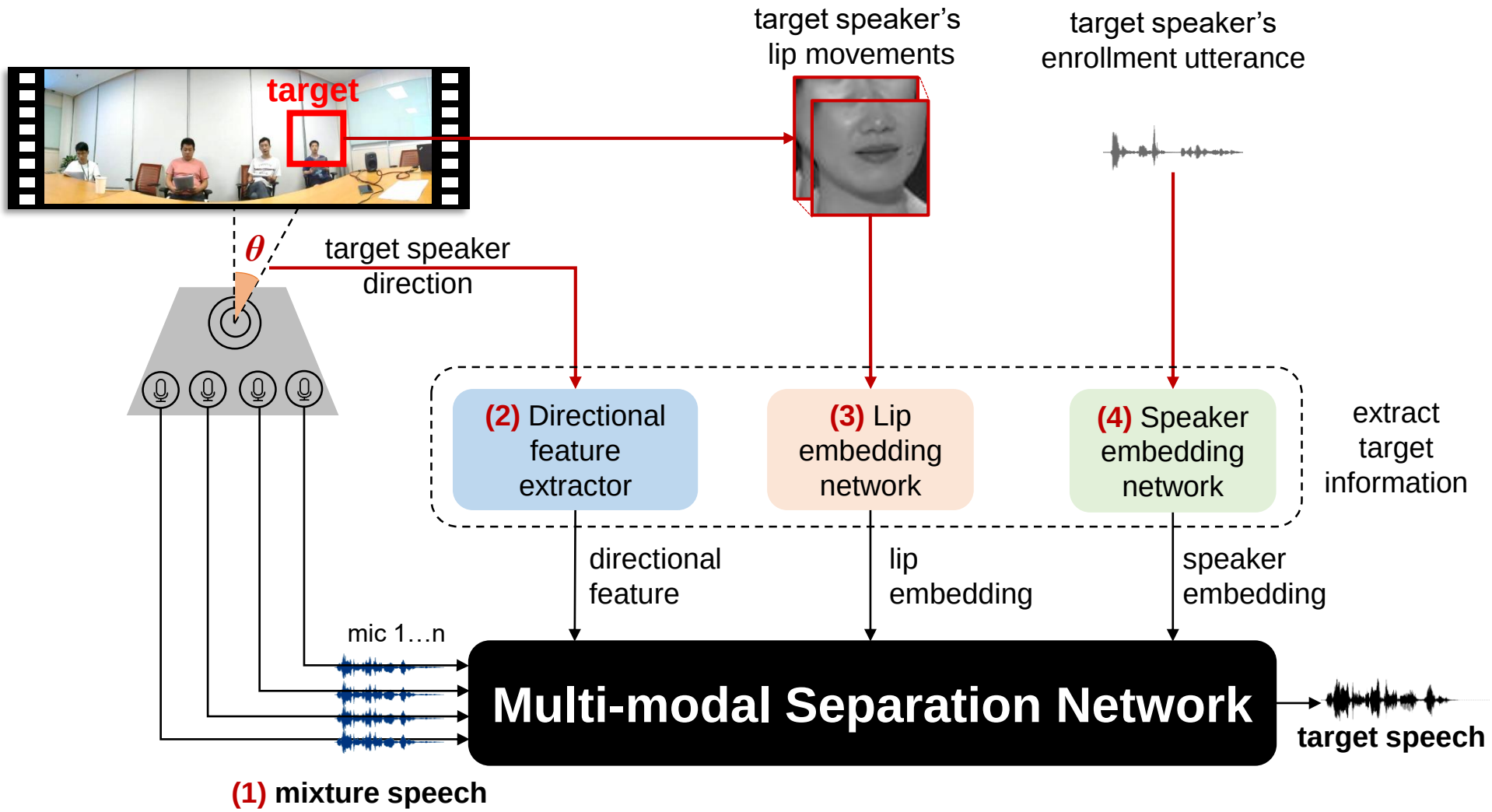


Enhanced voice filter

Outline

- Mask-based -> Filter-based
- Blind Source Separation -> Target Speaker Extraction
- **Single Modality -> Multi-Modality**
- Summary

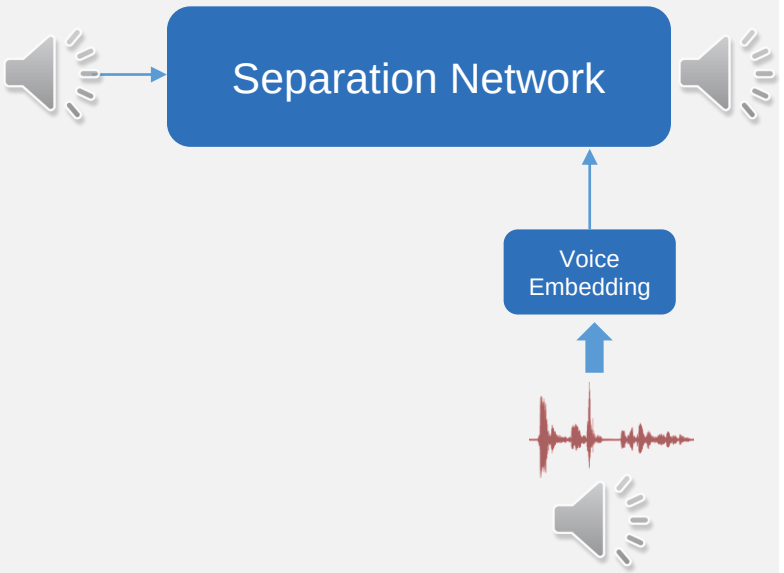
Multi-modal Target Speech Separation



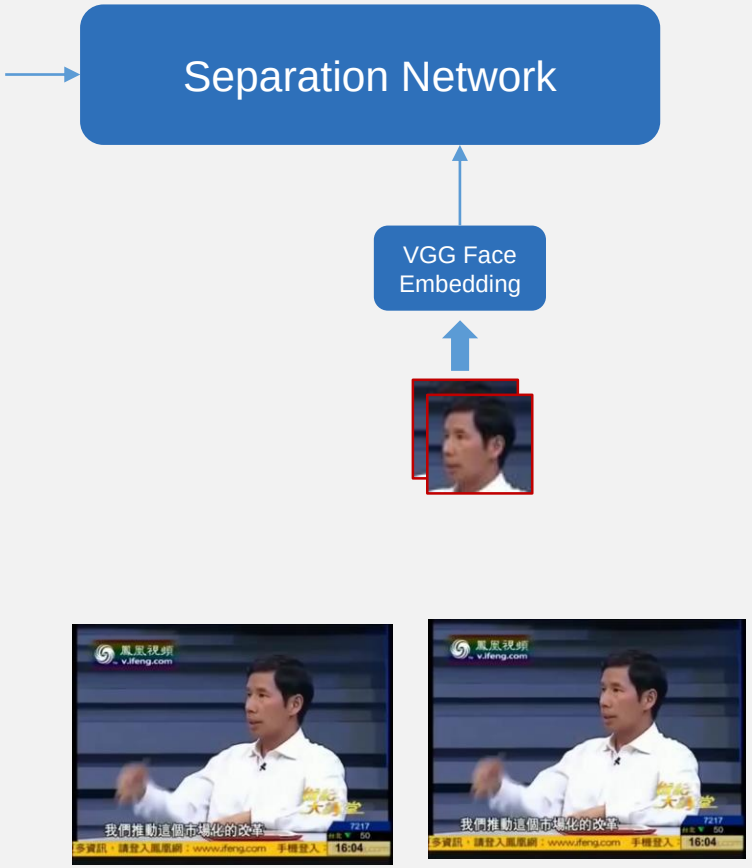
Multi-modal Target Speech Separation

Tencent AI Lab

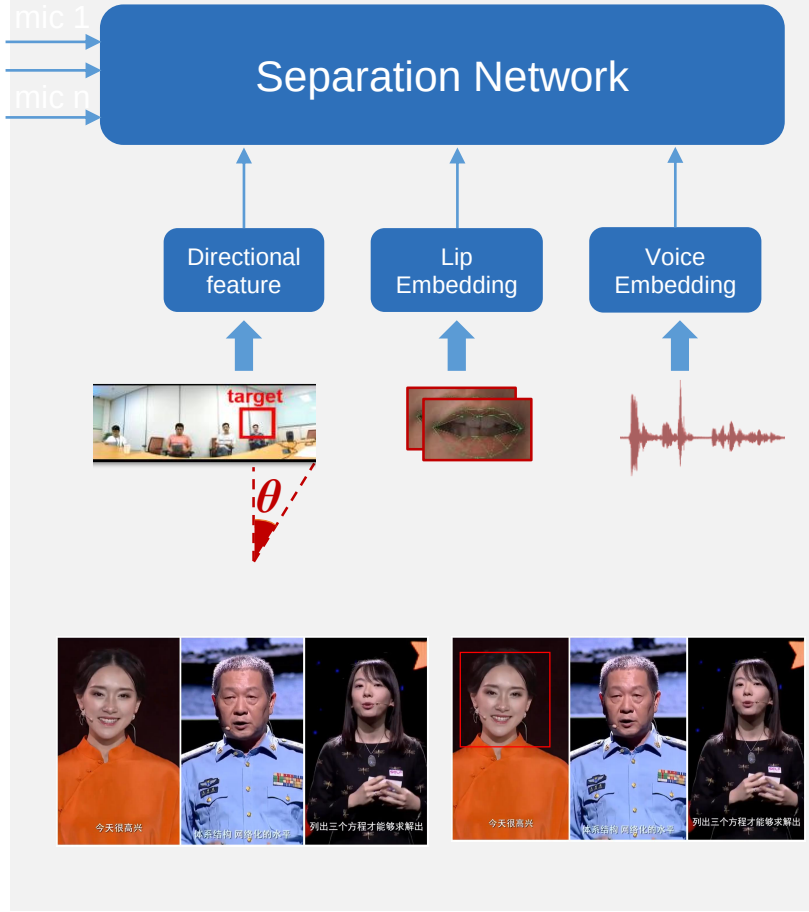
Voice Filter



Looking to Listen



Multi-modal



Multi-modal Speech Separation: Front Face

Mixed Speech

Generated for a simulated acoustic environment

Speech Extracted for a Target Speaker

using target speaker's direction, target speaker's mouth movement, and phase difference captured by microphone array

Multi-modal Speech Separation: Lateral Face

Mixture

Enhanced

More demos: https://jupiterethan.github.io/av-enh.github.io/

Impact of Different Modalities for Separation/Enhancement

Tencent AI Lab

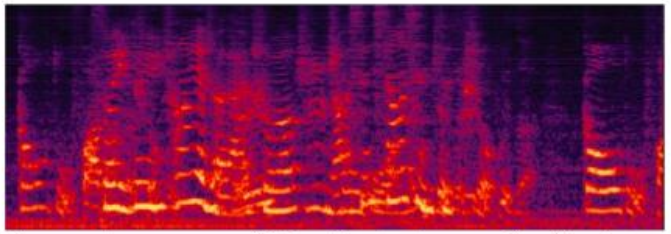
Input Modalities					
Directional Feature	Enrolled Voice	Target Lip	SDR (dB)	PESQ	WER (%)
√			16.9	3.24	11.3
	√		14.8	2.98	14.7
		√	16.6	3.01	19.6
√	√		17.1	3.25	10.5
√	√	√	**17.6**	3.28	10.0
√		√	**17.5**	3.28	10.3

Difference from **speaker embedding** is small。

We can stick to use only Directional feature + Lip if speaker embedding not available

Better Robustness with Multiple Modalities

(a) video frames of target speaker's lip region

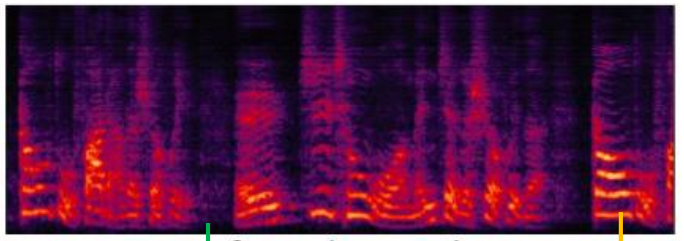


(b) spectrogram of the first channel of mixture

Directional Feature only

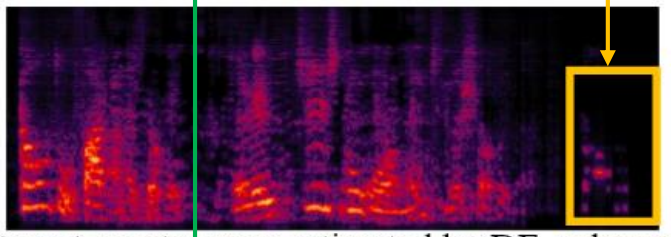
Lip Feature only

Multi-modal



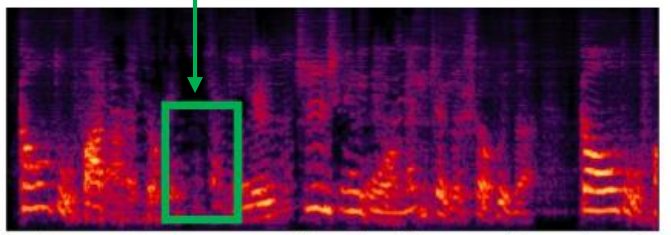
(c) spectrogram of reverberant clean target speech

Ground Truth



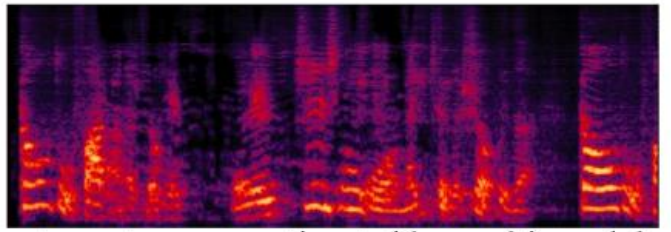
(d) target spectrogram estimated by DF-only model

over-suppresses the target speech when the speaker slightly turns his face



(e) target spectrogram estimated by audio-visual model

Leakage from interference when the speaker opens his mouth without speaking



(f) target spectrogram estimated by multi-modal model

No obvious issue

Robustness When One Modality is Missing or Inaccurate

Tencent AI Lab

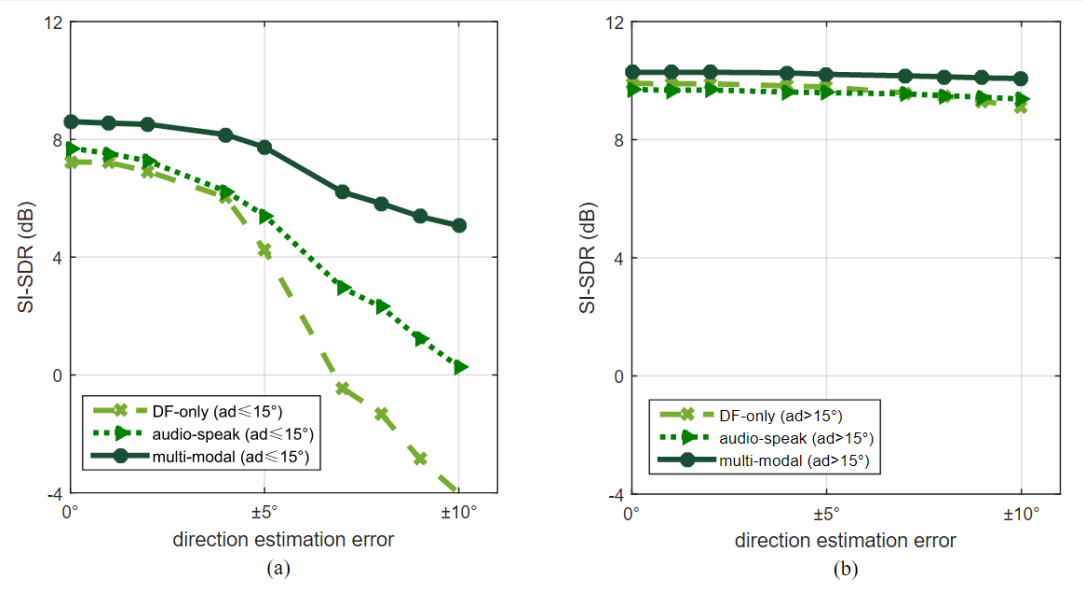
When Target Lip is Unavailable x% of Times

Lip Dropout Ratio	SI-SDR
0%	17.2
10%	17.1
20%	17.1
50%	17.0

Not important when the directional feature is accurate

Important when speakers are close or the target speaker is not facing the camera

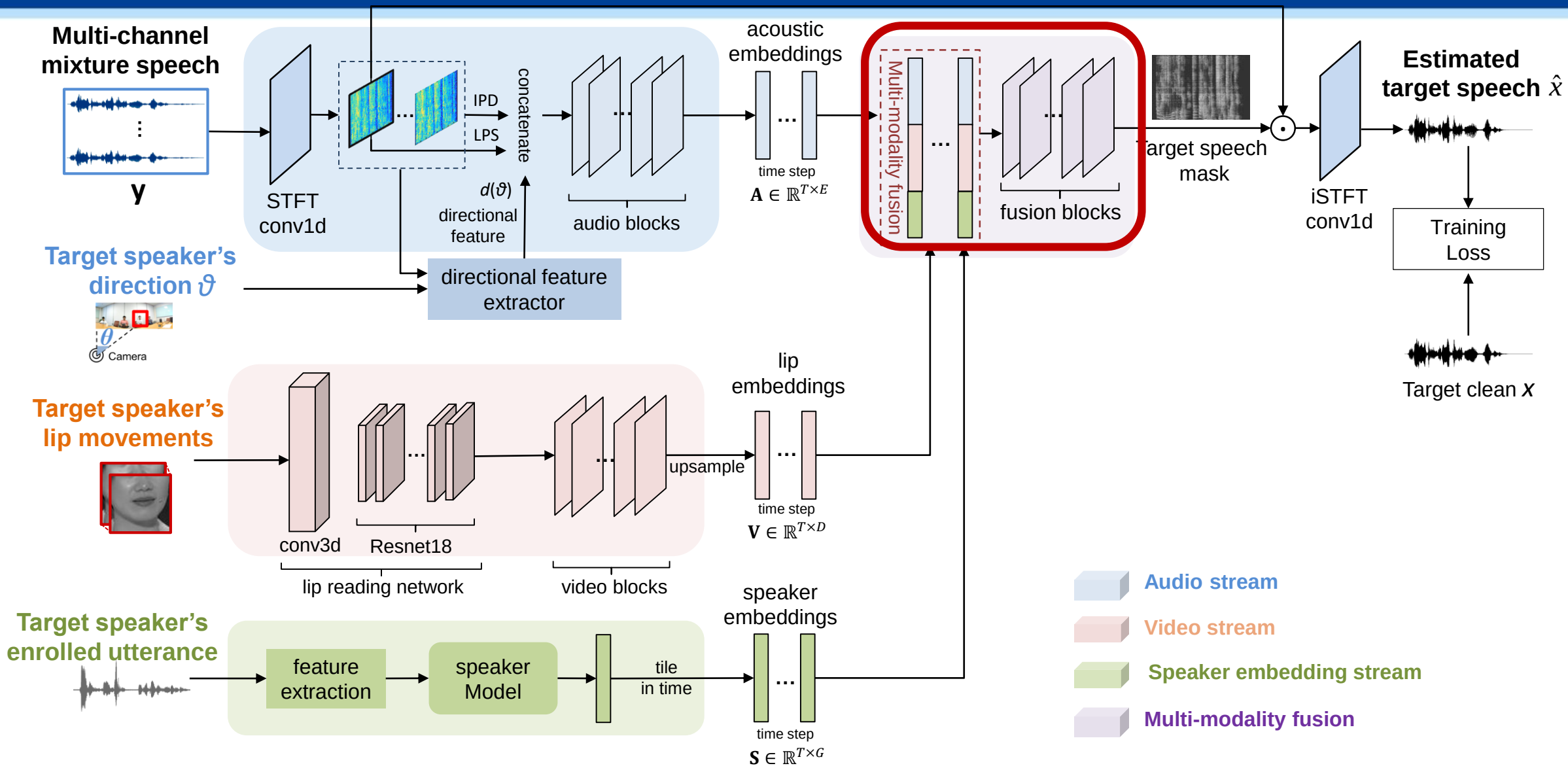
When Directional Feature is Incorrect



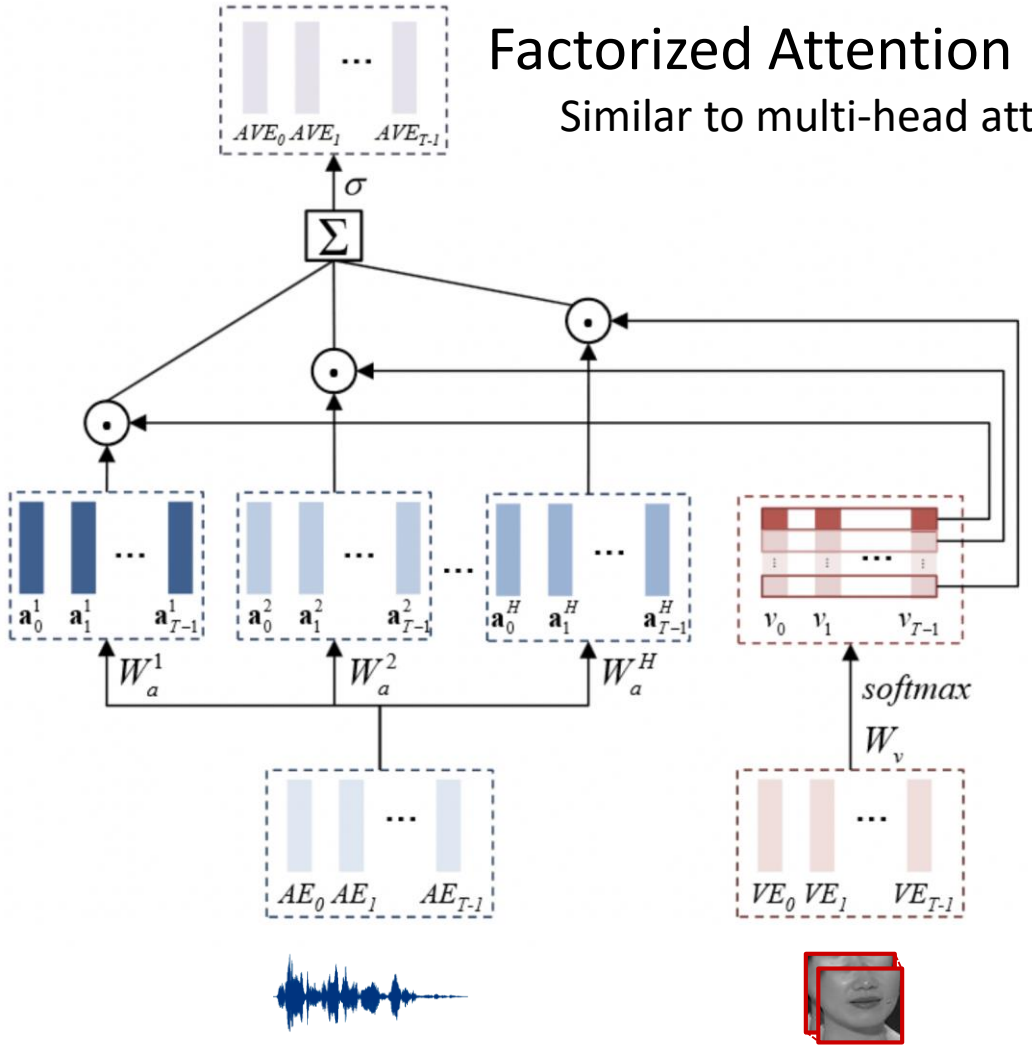
small angles (<15°) between target and interfering speakers

large angles (>=15°) between target and interfering speakers

Network Structure for Multimodal Fusion



Network Structure for Multimodal Fusion



Factorized Attention Fusion

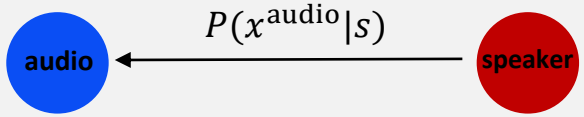
Similar to multi-head attention

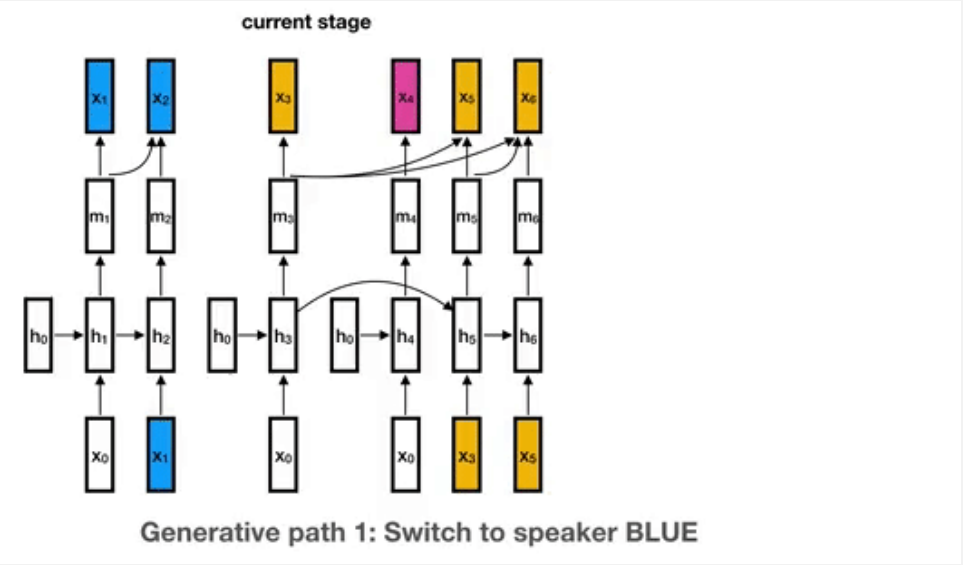
Fusion Method	SI-SDR (dB) avg	< 15°	> 90°
Concatenation	9.1	7.5	11.4
Rule-based Attention	8.9	6.3	11.9
Factorized Attention	**9.3**	7.6	11.5

Diarization

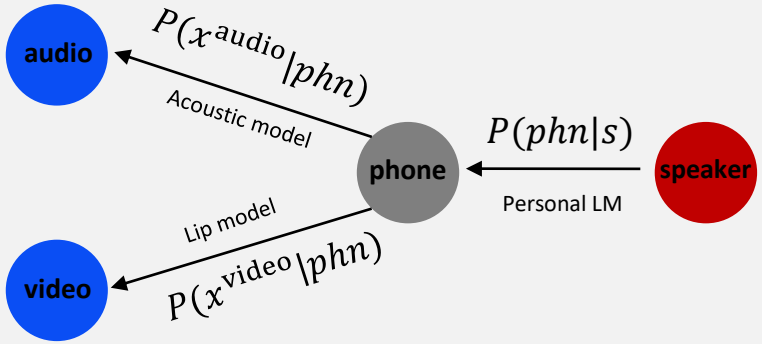
Audio-only

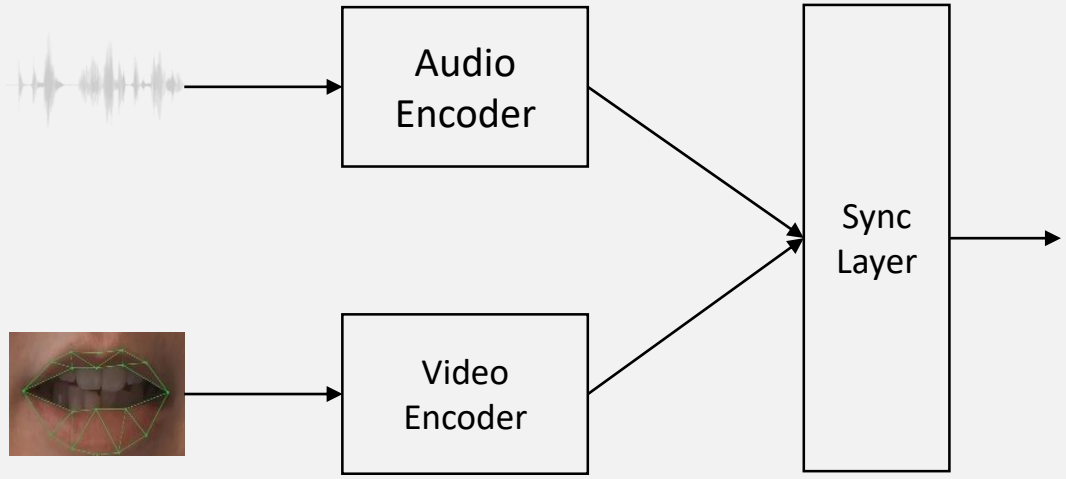
State-of-the-art: UIS-RNN [1]





Multi-modal

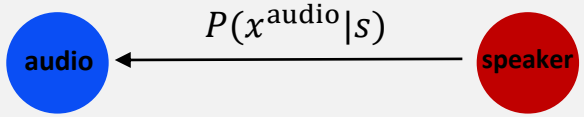




Diarization

Audio-only

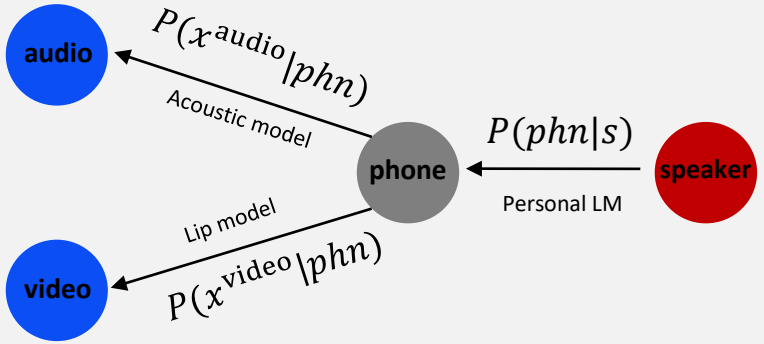
State-of-the-art: UIS-RNN [1]



Issues

1. unbounded number of speakers
2. not robust when two speakers' voices are similar
3. require labelled training data
4. cannot link the state to speaker id without pre-enrollment

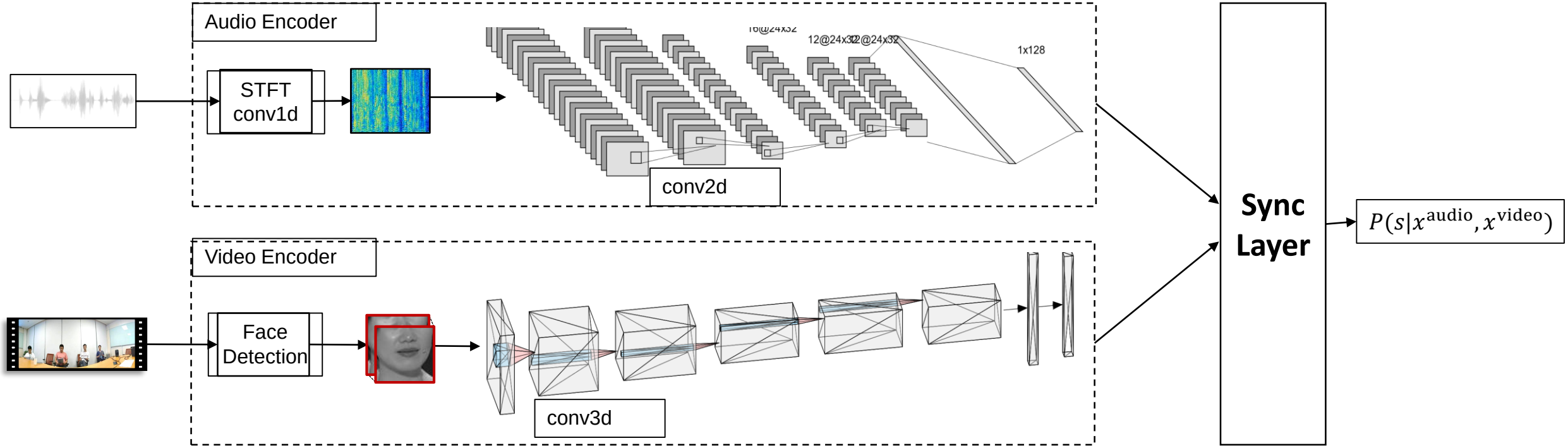
Multi-modal



Improvement

1. number of speakers are bounded by faces in the video
2. robust to speaker voice similarity
3. self-supervised learning with proposed dynamic triplet loss
4. can link speaker with his face

Multi-modal Diarization



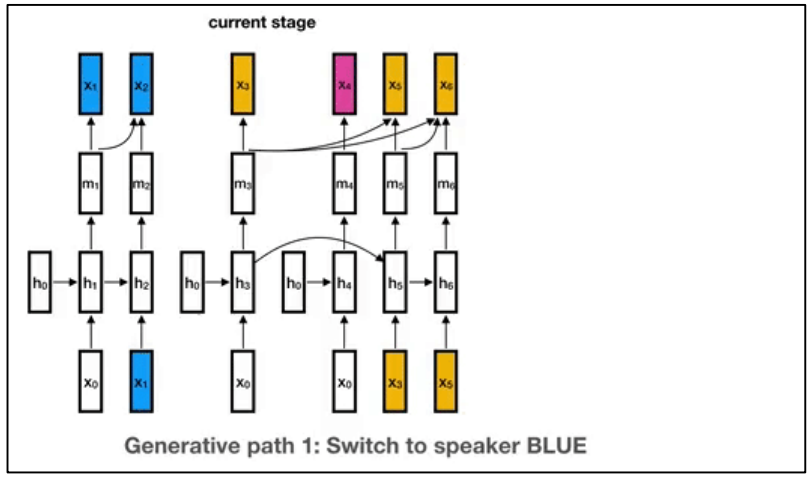
Sync layer is critical because most audio and video are recorded unsynchronizely due to the hardware and the speed difference between light wave and sound wave. For example, speaker at 3m away will cause 1 frame delay between audio and video.

find the best likely matched frame between audio and video. jointly trained

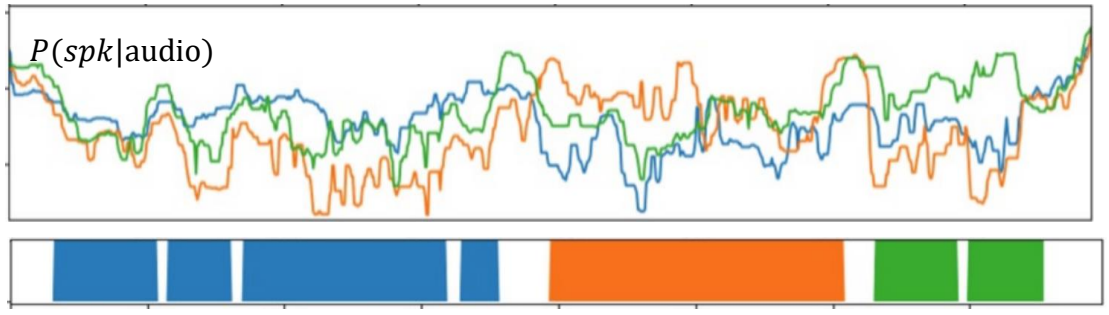
Multi-modal Diarization

Tencent AI Lab

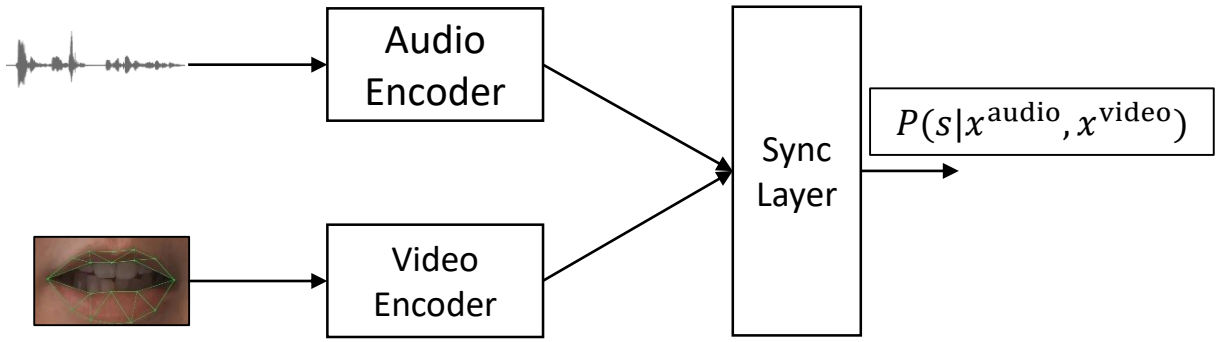
Audio-only



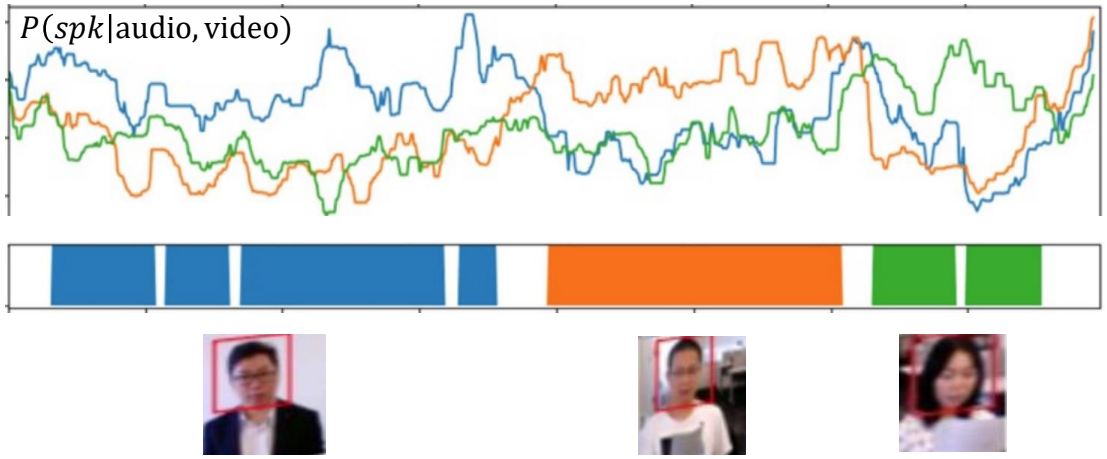
State-of-the-art: UIS-RNN
DER=25.6%



Multi-modal



Our multi-modal diarization
DER=17.0%



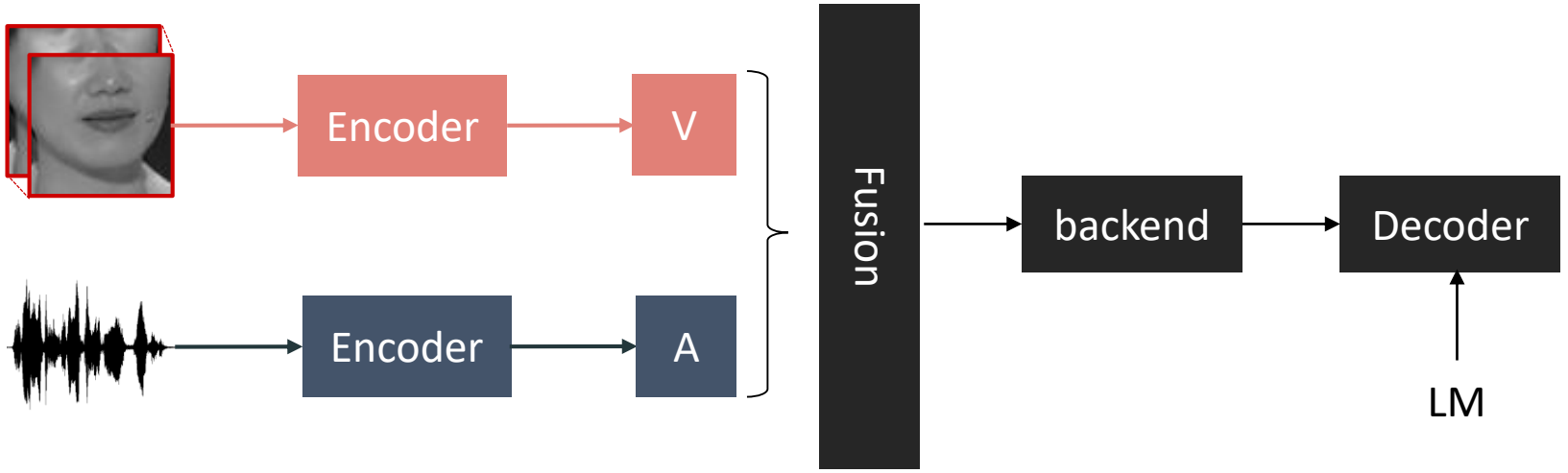
Multi-modal Tracking and Diarization (hard case)

Tencent AI Lab

Red Box: speaking

White Box: not speaking

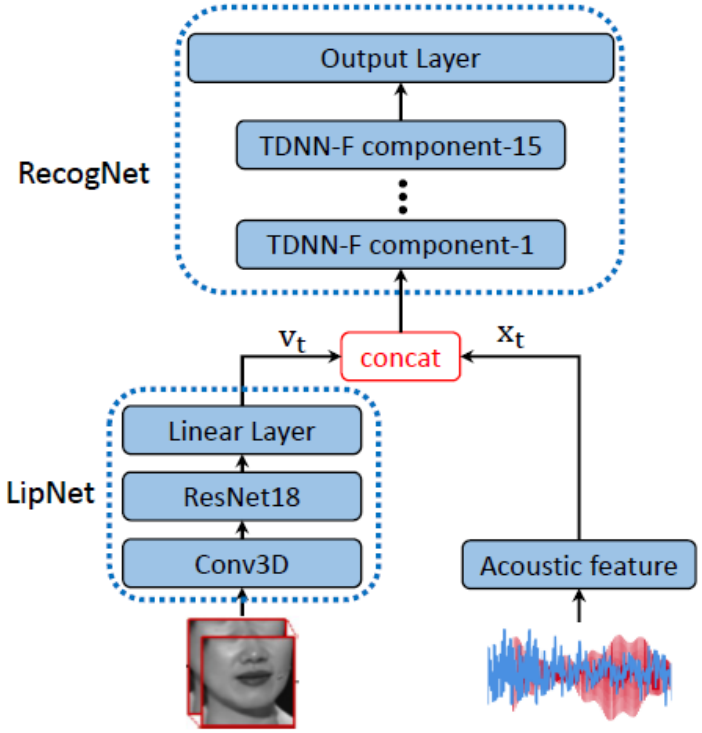
Multi-modal ASR on LRS2 Corpus (Dry Clean)



Models	V	A	A+V
TM-CTC [11]	65.0	15.3	13.7
TM-Seq2seq [11]	49.8	10.5	9.4
CTC/Attention [30]	63.5	8.3	7.0
LF-MMI TDNN	**48.86**	**6.71**	**5.3**

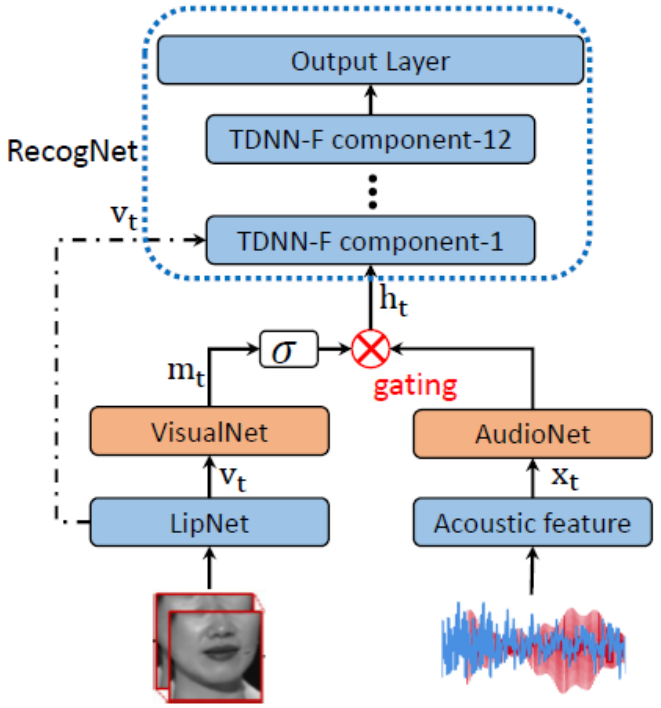
our system performs the best under all conditions on "dry clean"

Modality Fusion for ASR on Overlapped Speech



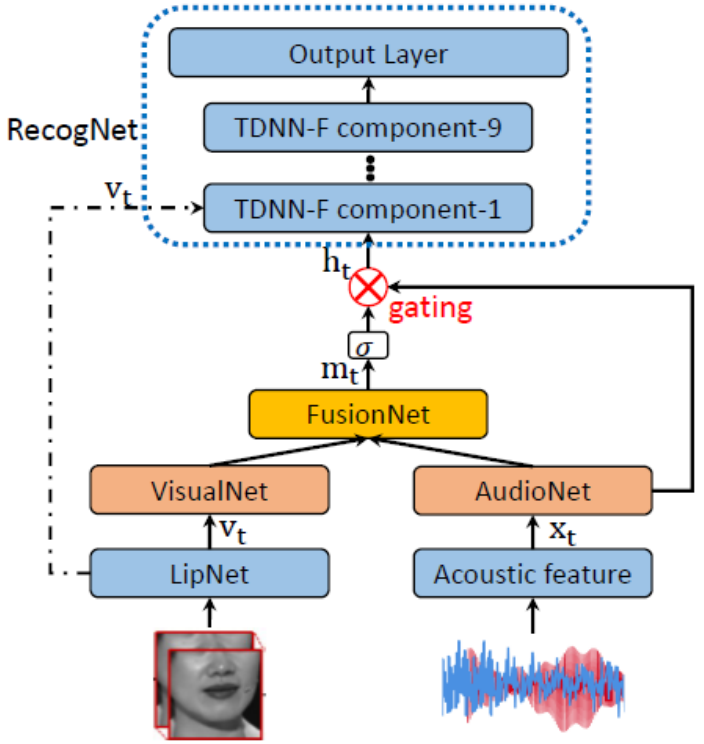
(a) Feature concatenation

WER=15.2%



(b) Visual modality driven gate

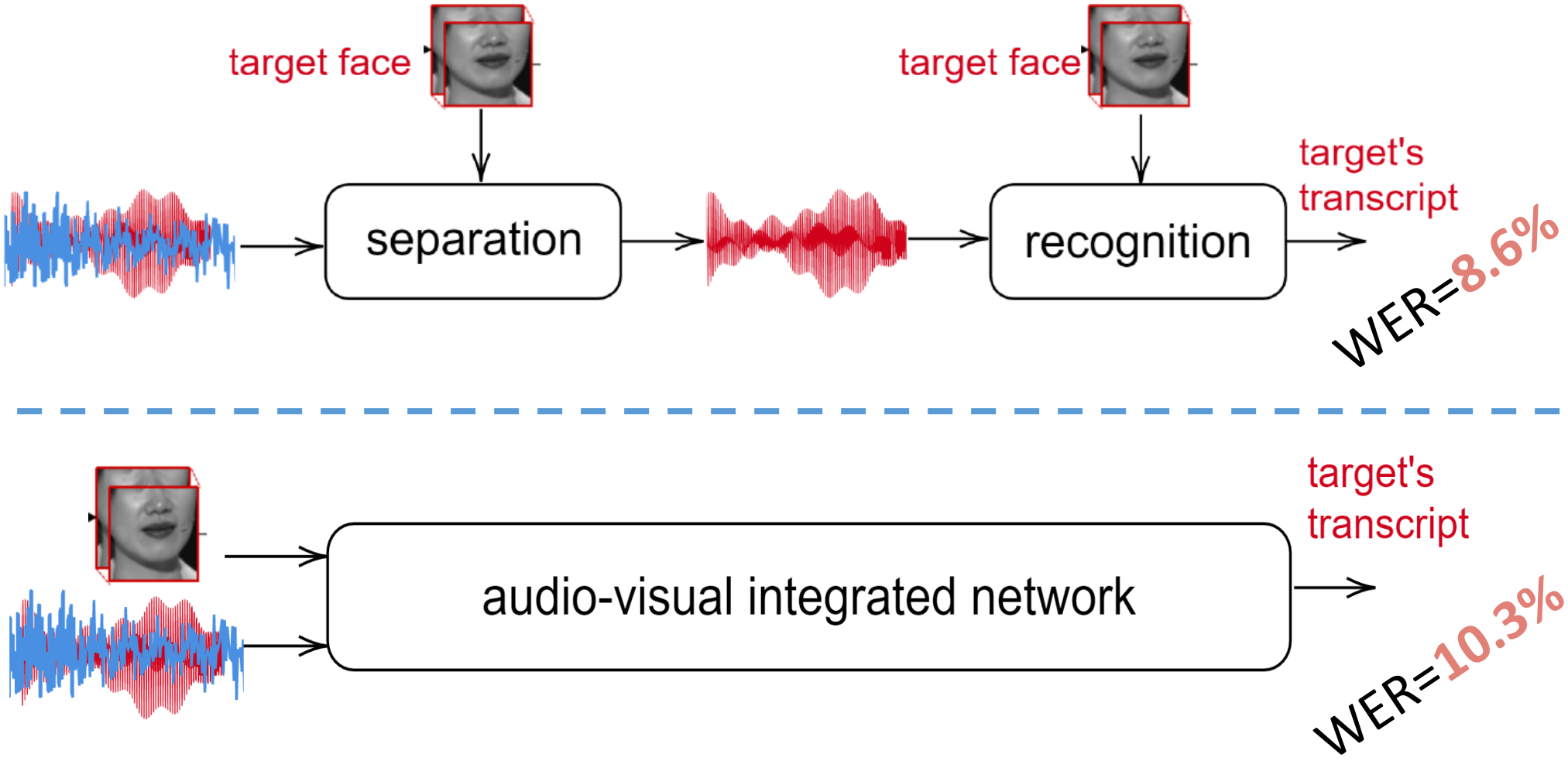
WER=11.5%



(c) Audio-visual modality driven gate

WER=10.3%

Multi-modal ASR for Overlapped Speech



Tracking+Diarization+Separation+ASR (Easy Case)

Tencent AI Lab

Raw input

Red box: speaking

Trace 

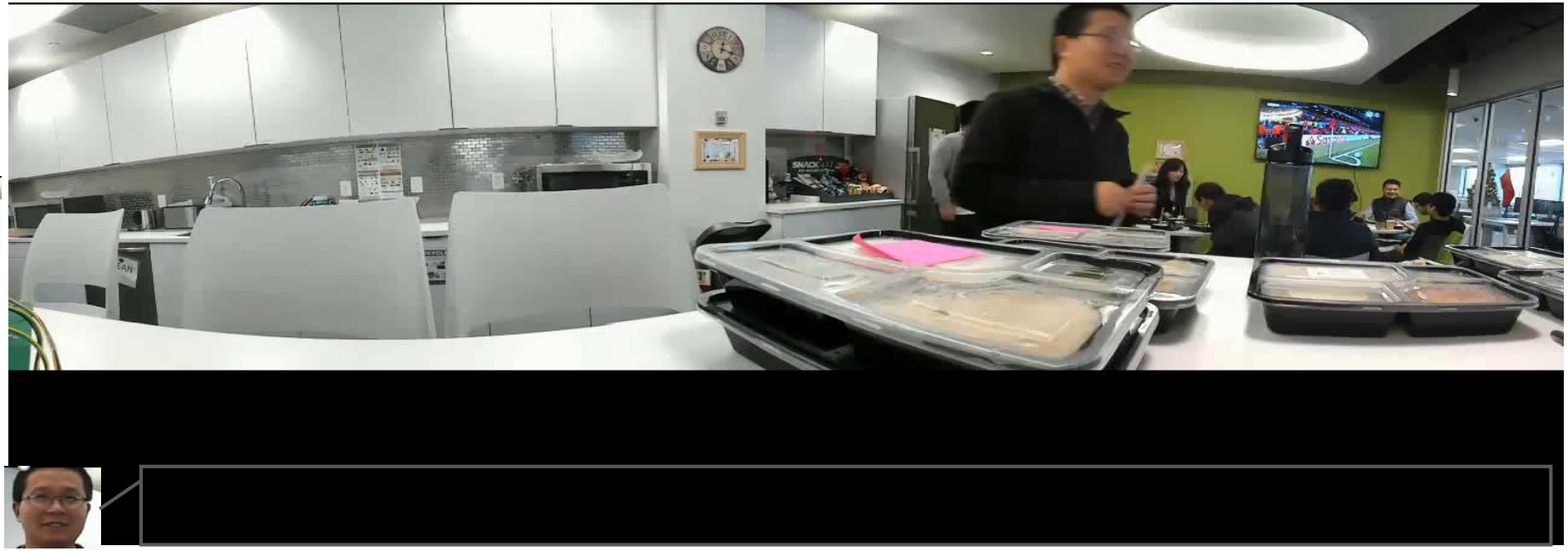
White box: not speaking

Tracking+Diarization+Separation+ASR (Hard Case)

Tencent AI Lab

Raw input

Trace 



Outline

- ❑ Mask-based -> Filter-based
- ❑ Blind Source Separation -> Target Speaker Extraction
- ❑ Single Modality -> Multi-Modality
- ❑ **Summary**

Summary

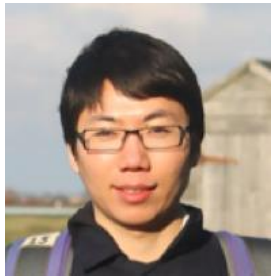
- Mask-based -> Filter-based

$$S_{t,f} = \boldsymbol{Mask}_{t,f} \times Y_{t,f} \quad \text{->} \quad S_{t,f} = \mathbf{W}_{t,f}^{H}\mathbf{Y}_{\mathrm{t,f}}$$

- Blind Source Separation -> Target Speaker Extraction
  - Exploit the information that can identify the speaker
- Single Modality -> Multi-Modality
  - Multi-modality is much more robust/better performing than single-modality
  - Better model of separation/recognition are yet to be found

Contributor Acknowledgement

Tencent AI Lab


Shi-Xiong Zhang


Yong XU


Meng YU


Chao WENG


Chunlei ZHANG


Jianming LIU


Max W. Y. Lam

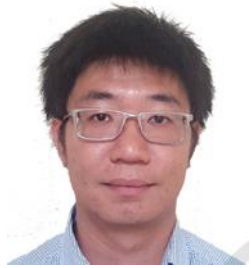

Lianwu CHEN


Xuan JI


Deyi TUO


Dan SU

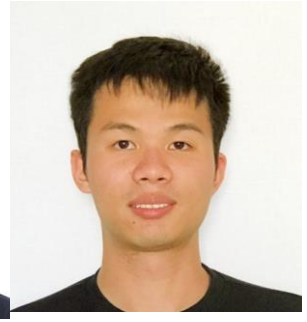

Jun WANG


Jimeng ZHENG


Jie CHEN


Rongzhi GU


Jian WU


Jianwei YU


Fahimeh BAHMANINEZHAD


Ke Tan


Aswin S. SUBRAMANIAN


Yifan DING

References

- multi-channel mask -> complex filter with end-to-end training
  - Yong Xu, Meng Yu, Shi-Xiong Zhang, Lianwu Chen, Chao Weng, Jianming Liu, Dong Yu, “Neural Spatio-Temporal Filtering for Target Speech Separation”, submitted to Interspeech 2020
  - Rongzhi Gu, Shi-Xiong Zhang, Lianwu Chen, Yong Xu, Meng Yu, Dan Su, Yuexian Zou, Dong Yu, "Enhancing End-To-End Multi-Channel Speech Separation via Spatial Feature Learning", ICASSP 2020
  - Yong Xu, Chao Weng, Like Hui, Jianming Liu, Meng Yu, Dan Su, Dong Yu, "Joint Training of Complex Ratio Mask Based Beamformer And Acoustic Model for Noise Robust ASR", ICASSP 2019
  - Rongzhi Gu, Jian Wu, Shi-Xiong Zhang, Lianwu Chen, Yong Xu, Meng Yu, Dan Su, Yuexian Zou, and Dong Yu. "End-to-end multi-channel speech separation." arXiv preprint arXiv:1905.06286 (2019).
- Multi-modal Diarization and Speech Separation/Extraction/Recognition
  - Rongzhi Gu, Shixiong Zhang, Yong Xu, Lianwu Chen, Yuexian Zou, Dong Yu, “Multi-modal Multi-channel Target Speech Separation”, IEEE Journal of Selected Topics in Signal Processing, 2020.
  - Ke Tan, Yong Xu, Shixiong Zhang, Meng Yu, Dong Yu, “Audio-Visual Speech Separation and Dereverberation with a Two-Stage Multimodal Network”, IEEE Journal of Selected Topics in Signal Processing, 2020
  - Jianwei Yu, Shixiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, Dong Yu, "Audio-Visual Recognition of Overlapped Speech for the LRS2 Dataset", ICASSP 2020
  - Yifan Ding, Yong Xu, Shi-Xiong Zhang, Yahuan Cong, and Liqiang Wang "Self-supervised learning for audio-visual speaker diarization." ICASSP 2020.
  - Jian Wu, Yong Xu, Shi-Xiong Zhang, Lianwu Chen, Meng Yu, Lei Xie, Dong Yu, "Time Domain Audio Visual Speech Separation", ASRU 2019

References

- Blind Separation -> Speaker/Text/Direction-Aware Speech Extraction
  - Meng Yu, Xuan Ji, Bo Wu, Dan Su, Dong Yu, “End-to-End Multi-Look Keyword Spotting”, submitted to Interspeech 2020
  - Xuan Ji, Meng Yu, Jie Chen, Jimeng Zheng, Dan Su, Dong Yu, "Integration of Multi-Look Beamformers for Multi-Channel Keyword Spotting", ICASSP 2020
  - Max W. Y. Lam, Jun Wang, Dan Su, Dong Yu, "Mixup-Breakdown: A Consistency Training Method for Improving Generalization of Speech Separation Models", ICASSP 2020.
  - Xuan Ji, Meng Yu, Chunlei Zhang, Dan Su, Tao Yu, Xiaoyu Liu, Dong Yu, "Speaker-Aware Target Speaker Enhancement by Jointly Learning with Speaker Embedding Extraction", ICASSP 2020.
  - Aswin Shanmugam Subramanian, Chao Weng, Meng Yu, Shi-Xiong Zhang, Yong Xu, Shinji Watanabe, Dong Yu, "Far-Field Location Guided Target Speech Extraction Using End-To-End Speech Recognition Objectives", ICASSP 2020
  - Fahimeh Bahmaninezhad, Shi-Xiong Zhang, Yong Xu, Meng Yu, John HL Hansen, and Dong Yu. "A Unified Framework for Speech Separation." in submission to Speech Communications (2019).
  - Rongzhi Gu, Lianwu Chen, Shixiong Zhang, Jimeng Zheng, Meng Yu, Yong Xu, Dan Su, Yuexian Zou and Dong Yu, “Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information”, Interspeech 2019
  - Fahimeh Bahmaninezhad, Jian Wu, Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Meng Yu and Dong Yu, “A comprehensive study of speech separation: spectrogram vs waveform separation”, Interspeech 2019
  - Meng Yu, Xuan Ji, Yi Gao, Lianwu Chen, Jie Chen, Jimeng Zheng, Dan Su, Dong Yu, "Text-Dependent Speech Enhancement for Small-Footprint Robust Keyword Detection", Interspeech 2018.
  - Jun Wang, Jie Chen, Dan Su, Lianwu Chen, Meng Yu, Yanmin Qian, Dong Yu, "Deep Extractor Network for Target Speaker Recovery From Single Channel Speech Mixtures", Interspeech 2018

References

Tencent AI Lab

❑ Background papers / prior arts

- ❖ Jacob Benesty, Jingdong Chen, and Emanuël AP Habets, “multi-channel speech enhancement with filters “, ch5 in ”Speech enhancement in the STFT domain”, Springer Science & Business Media, 2011.
- ❖ Yuxuan Wang, Arun Narayanan, and DeLiang Wang. "On training targets for supervised speech separation." IEEE/ACM transactions on audio, speech, and language processing, 2014.
- ❖ Williamson, Donald S., Yuxuan Wang, and DeLiang Wang. "Complex ratio masking for monaural speech separation." IEEE/ACM transactions on audio, speech, and language processing, 2015
- ❖ John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. "Deep clustering: Discriminative embeddings for segmentation and separation." ICASSP 2016.
- ❖ Dong Yu, Morten Kolbæk, Zheng-Hua Tan, Jesper Jensen, “Permutation Invariant Training of Deep Models for Speaker-Independent Multi-Talker Speech Separation”, ICASSP 2017
- ❖ Morten Kolbæk, Dong Yu, Zheng-Hua Tan, Jesper Jensen, “Multi-talker Speech Separation with Utterance-level Permutation Invariant Training of Deep Recurrent Neural Networks”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017
- ❖ Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation." ACM Transactions on Graphics, 2018.
- ❖ Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, and Ignacio Lopez Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking", Interspeech 2019
- ❖ Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang "Fully supervised speaker diarization." ICASSP 2019

Tencent
AI Lab

Thank You!