Advances in end-to-end neural source separation

YI LUO (罗艺)

NEURAL ACOUSTIC PROCESSING LABORATORY (NAPLAB)

COLUMBIA UNIVERSITY

Prerequisite

If you are not that familiar with source separation:

- Wang, DeLiang, and Jitong Chen. "Supervised speech separation based on deep learning: An overview." IEEE/ACM Transactions on Audio, Speech, and Language Processing 26.10 (2018): 1702-1726.
- Rafii, Zafar, et al. "An overview of lead and accompaniment separation in music." IEEE/ACM Transactions on Audio, Speech, and Language Processing 26.8 (2018): 1307-1335.
- Gannot, Sharon, et al. "A consolidated perspective on multimicrophone speech enhancement and source separation." IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.4 (2017): 692-730.

Outline

End-to-end neural source separation: definition and progress

- Journey to the state-of-the-art

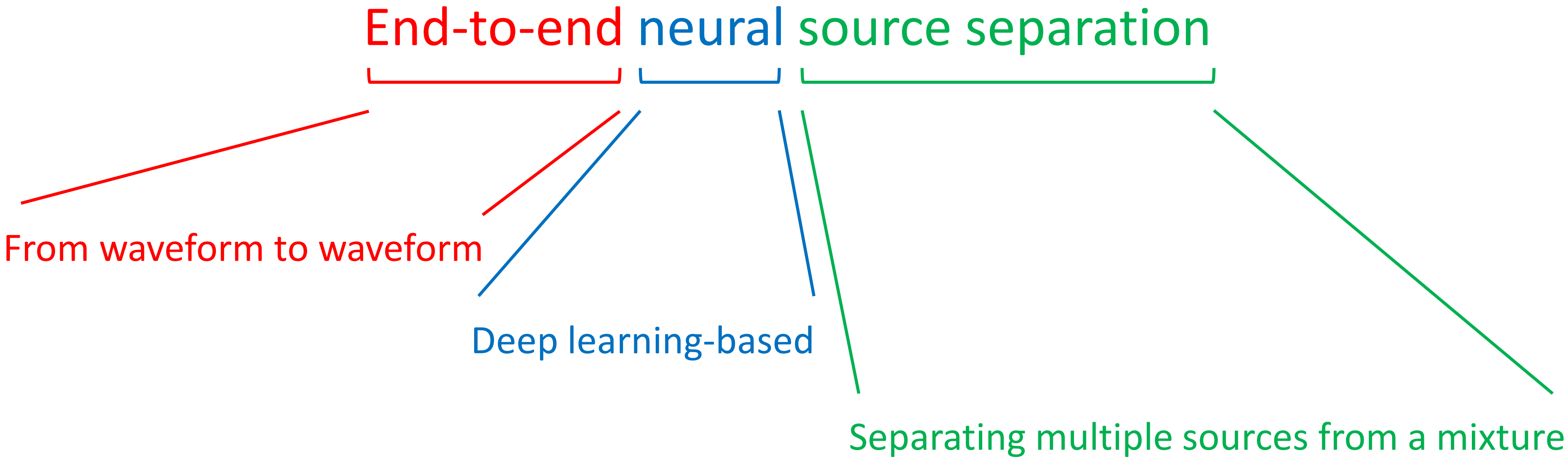
Single-channel systems

- Explicit replacement of short-time Fourier transform (STFT)
- Waveform-to-waveform mapping

Multi-channel systems

- Neural beamformers
- Single-channel extensions

Challenges and future directions



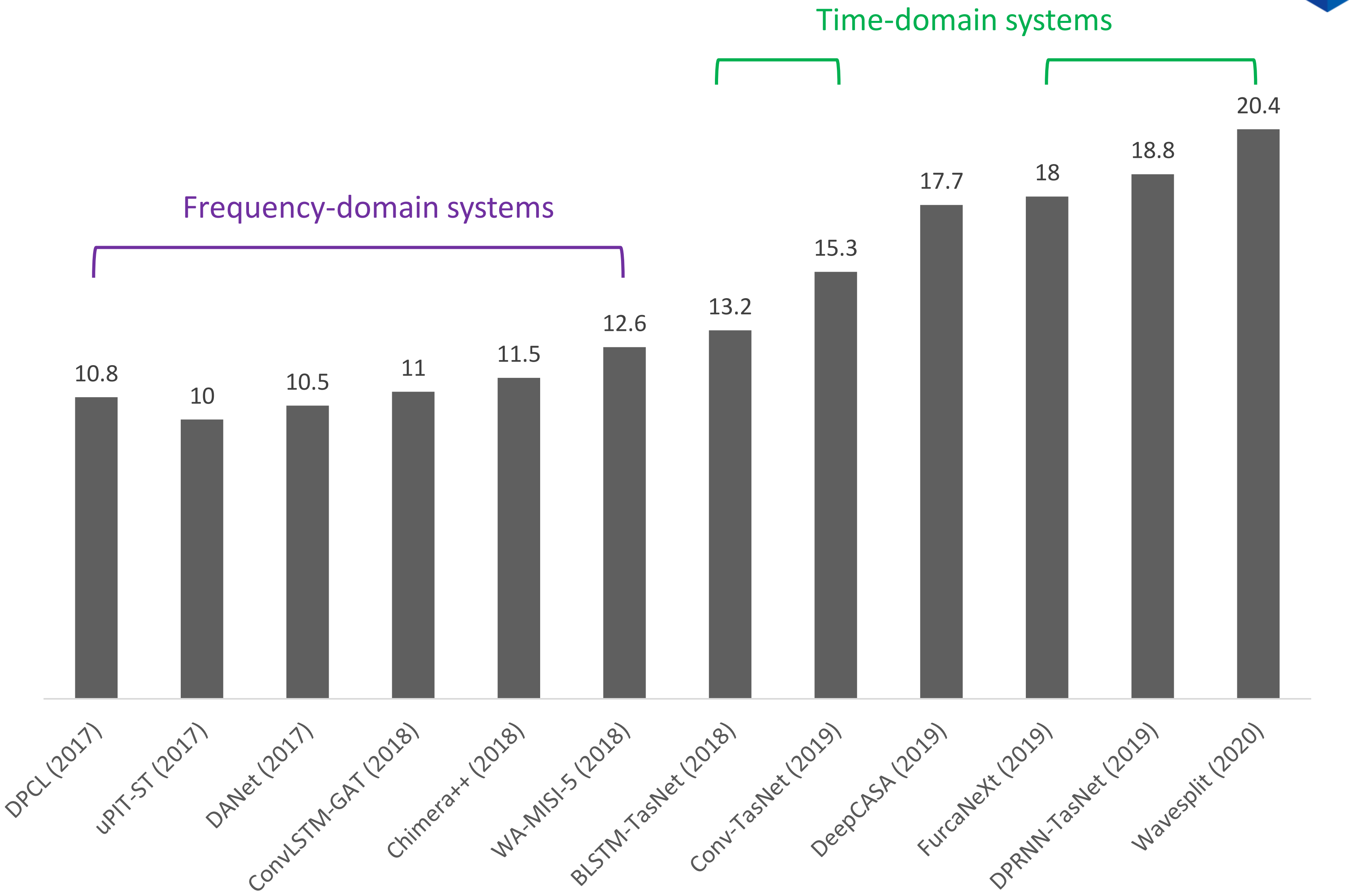
End-to-end neural source separation

From waveform to waveform

Deep learning-based

Separating multiple sources from a mixture

SI-SDR improvement (dB) on WSJ0-2mix dataset



Outline

End-to-end neural source separation: definition and progress

- Journey to the state-of-the-art

Single-channel systems

- Explicit replacement of short-time Fourier transform (STFT)
- Waveform-to-waveform mapping

Multi-channel systems

- Neural beamformers
- Single-channel extensions

Challenges and future directions

Source separation: problem definition

Observation: $y_i = \sum_{j=1}^{C} x_i^j + n_i, i = 1, \dots, K$

- y_i: mixture signal at i-th microphone
- x_i^j: j-th source at i-th microphone
- n_i: additional noise at i-th microphone
- In reverberant environments, x_i^j contains reflections of the direct path signal (reverberations)

Target: x_1^j

- All sources at the first (reference) microphone

Single-channel systems: frequency-domain

Single-channel separation:

- Observation: $y = \sum_{j=1}^{C} x^j + n \ (K = 1)$
- Target: x^j

Conventional (mainstream) frequency-domain method: time-frequency masking (T-F masking)

- Map the observation mixture into time-frequency domain (Short-time Fourier transform)
- Assumption: each time-frequency bin (T-F bin) can be assigned/classified to one of the sources
- Hard/soft assignment -> hard/soft classification problem
- T-F masks: classification assignments/probabilities

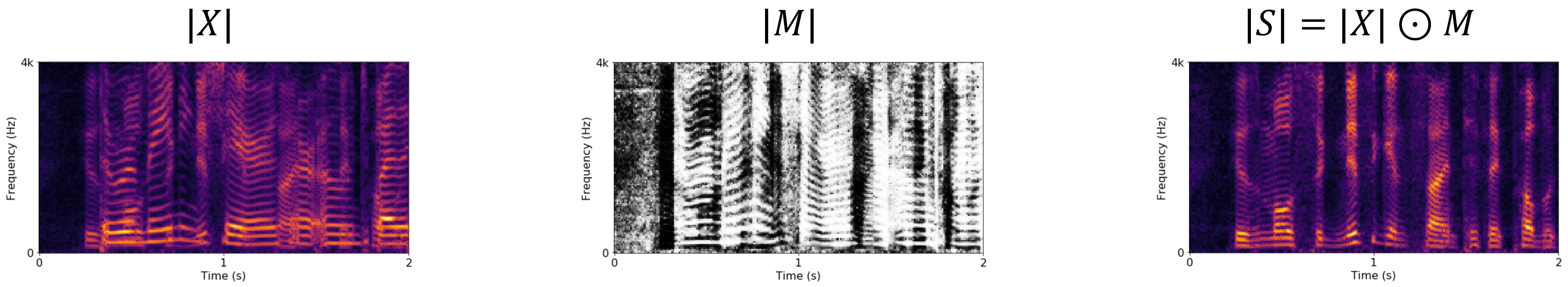
Oracle (Ideal) masks: calculated from the relative energies of the clean sources

- Ideal binary mask (IBM), ideal ratio mask (IRM), etc.
- Typically used as the training targets in supervised training settings

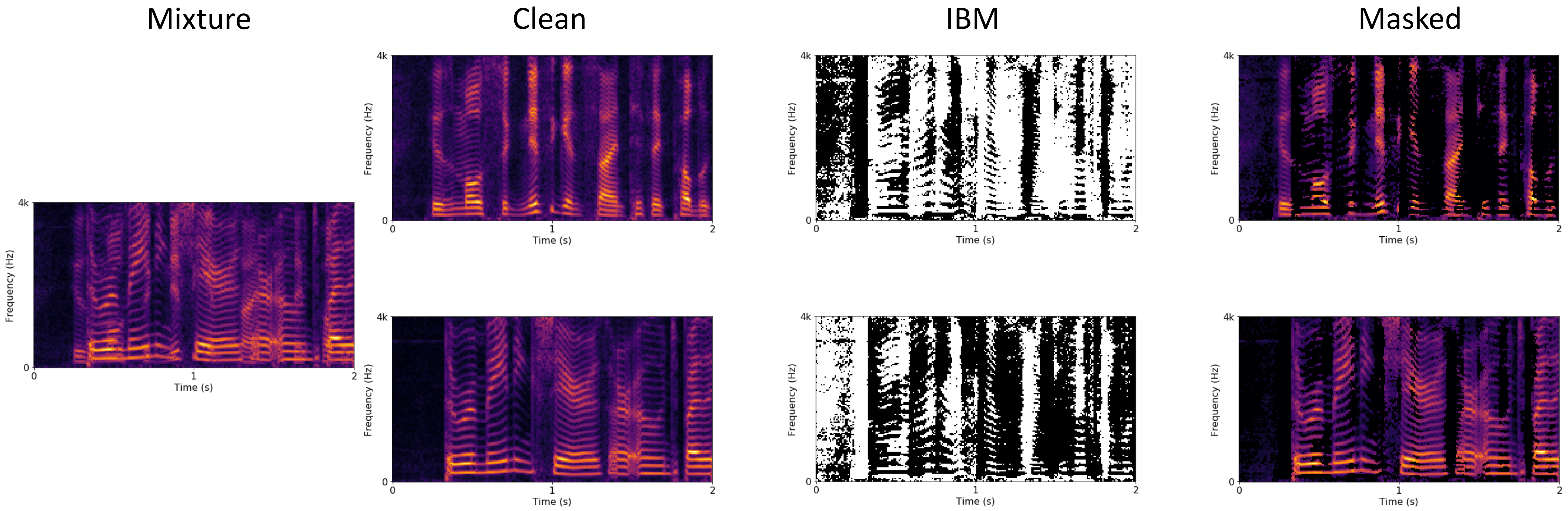
Single-channel systems: frequency-domain

An **encoder-separator-decoder** formulation:

- Encoder: STFT (spectrograms), $x \rightarrow |X|$ (1D -> 2D)
- Separator: T-F mask estimation, $|X| \rightarrow \mathrm{M}$, $|S| = |X| \odot M$
- Decoder: inverse STFT, $|S| \rightarrow \hat{x}$



Single-channel systems: frequency-domain



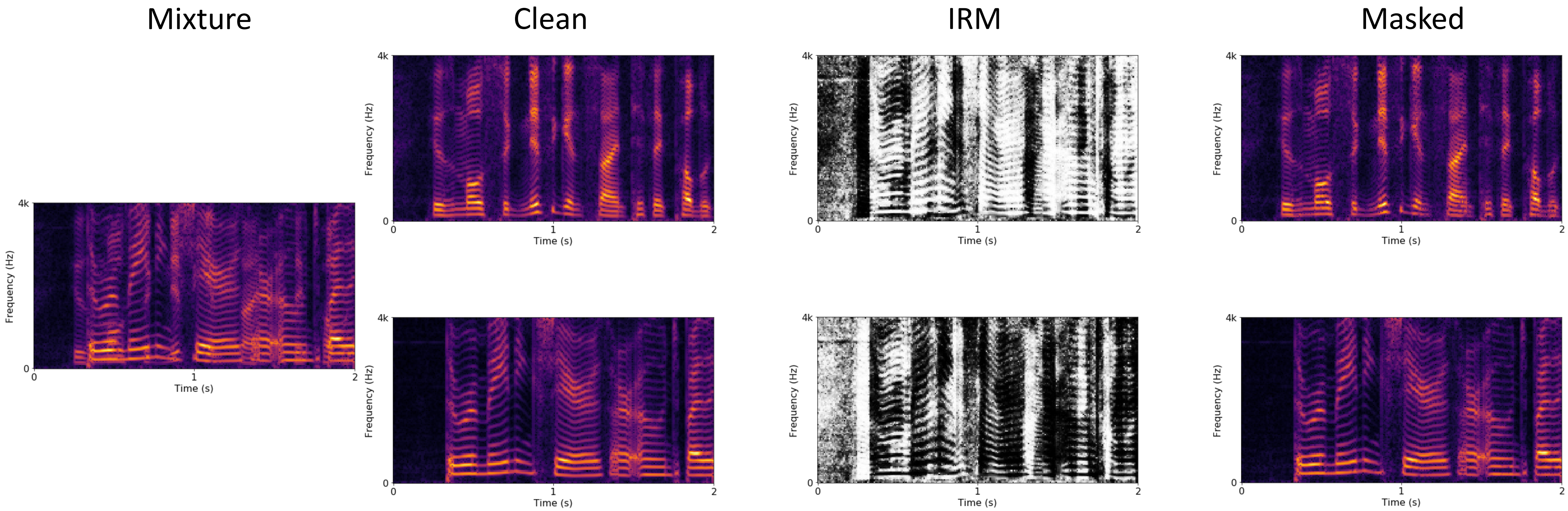
Mixture

Clean

IBM

Masked

Single-channel systems: frequency-domain



Mixture

Clean

IRM

Masked

Single-channel systems: time-domain

Drawbacks for T-F masking:

- Phase information is not fully utilized (except for certain types of T-F masks)
- Performance is upper-bounded by the oracle masks used as training targets
- Spectrograms (output of STFT) may not be optimal features for the task of separation

Time-domain systems: two directions

- Replace STFT with a real-valued (and trainable) module (Fourier basis as linear matrix multiplication)
- Directly learn a mapping between waveforms without an explicit feature transformation

Single-channel systems: replacing STFT

Discrete Fourier transform (DFT) matrix: $X = Wx$, where x is the N-point input signal, X is the DFT of the signal, and

$$W = \frac{1}{\sqrt{N}}\begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \omega^3 & \cdots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \omega^6 & \cdots & \omega^{2(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \omega^{3(N-1)} & \cdots & \omega^{(N-1)(N-1)} \end{bmatrix}$$

is the N×N complex-valued DFT matrix with $\omega = e^{-\frac{2\pi i}{N}}$

Replace W with a real-valued, trainable matrix (no need to be symmetric or square): $\hat{X} = \widehat{W}x$, with $\widehat{W} \in \mathbb{R}^{K\times N}$ a transformation matrix

Single-channel systems: replacing STFT

Fitting into the encoder-separator-decoder framework:

- Encoder: $\hat{X} = \widehat{W}x$ (compare with STFT)
- Separator: $S = \hat{X} \odot M$ (compare with T-F masking)
- Decoder: $\hat{x} = \hat{P}S$ (compare with inverse STFT)

Recall that for T-F masking:

- Encoder: STFT (spectrograms), $x \rightarrow X$ (1D -> 2D)
- Separator: T-F mask estimation, $X \rightarrow \mathrm{M}$, $S = X \odot M$
- Decoder: inverse STFT, $S \rightarrow \hat{x}$

Differences: encoder and decoder operations

Single-channel systems: objective/metric

Scale-invariant signal-to-distortion ratio (SI-SDR) [1]:

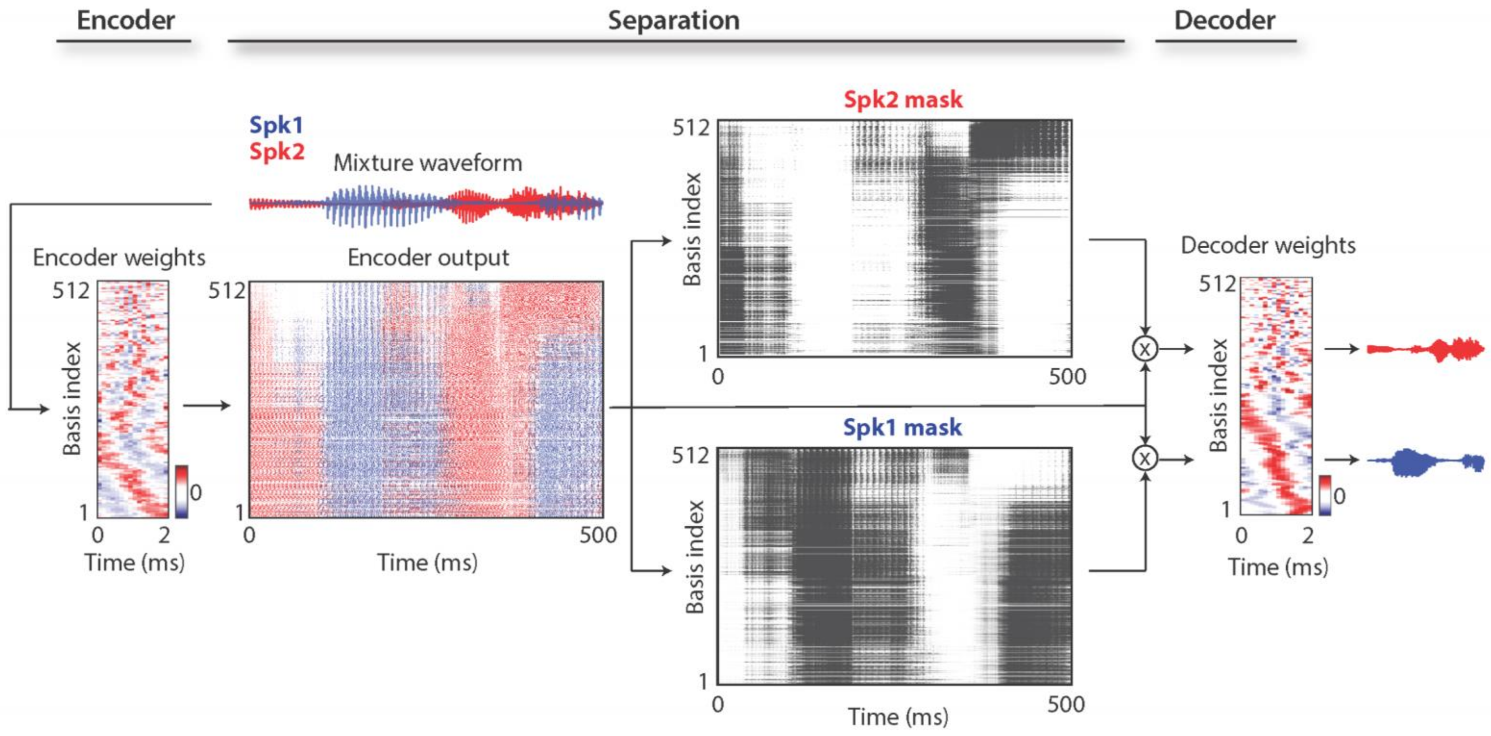
$$\begin{aligned}\text{SI-SDR} &= 10\log_{10}\left(\frac{||e_{\text{target}}||^2}{||e_{\text{res}}||^2}\right)\\ &= 10\log_{10}\left(\frac{||\frac{\hat{s}^T s}{||s||^2}s||^2}{||\frac{\hat{s}^T s}{||s||^2}s-\hat{s}||^2}\right)\end{aligned}$$

Other metrics: SNR/SDR/SIR/SAR…

Differentiable metrics can also be used as training objectives

Single-channel systems: TasNet

General model design: time-domain audio separation network (**TasNet**) [2-4]



TasNet: variants on the encoder/decoder

Going deeper: can we put constraints on the encoder/decoder?

DFT/inverse DFT can perfectly reconstruct the input

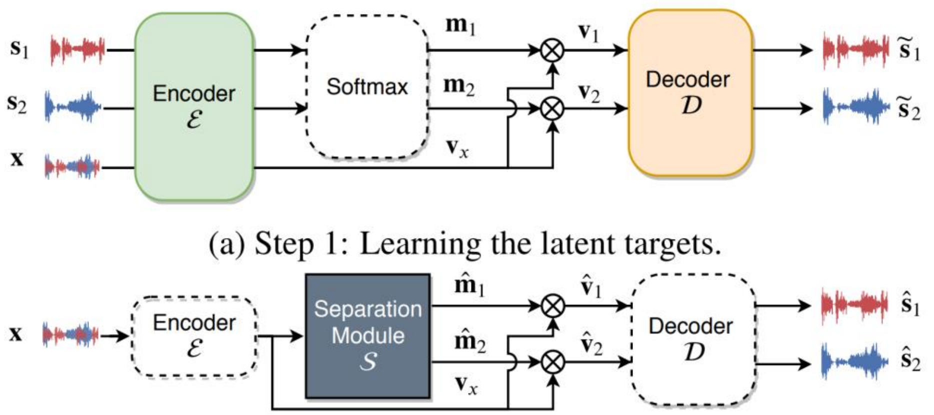
- In general, non-square $\widehat{W}$ and $\hat{P}$ (encoder/decoder matrices) cannot perfectly reconstruct the input

The representation should better reflect the target properties of the input

- For speech, a focus on lower frequency bands might be helpful

TasNet: variants on the encoder/decoder

Two-step separation [5]: first learn the linear encoder/decoder to reconstruct the input as good as possible, then freeze them during the training of the separator



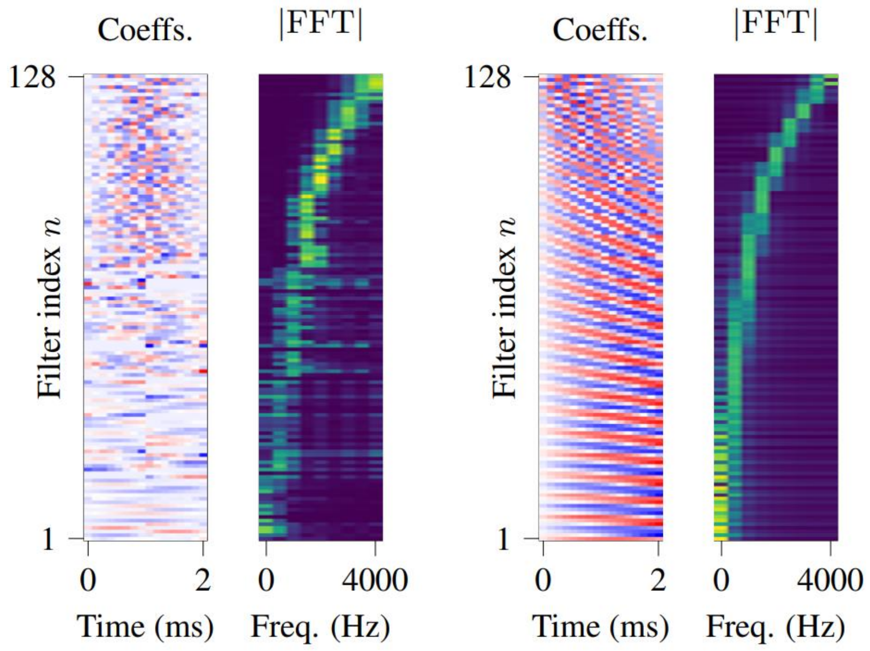
(a) Step 1: Learning the latent targets.

(b) Step 2: Training the separation module only.

Fig. 1: Training a separation network in two independent steps. For each step, the non-trainable parts are represented with a dashed line.

TasNet: variants on the encoder/decoder

Multi-phase Gammatone filterbank [6]: mimic the behavior of human auditory system



(a) Learned Filterbank (b) Proposed MP-GTF

TasNet: variants on objective

Wavesplit [7]: joint speaker identification and speaker separation

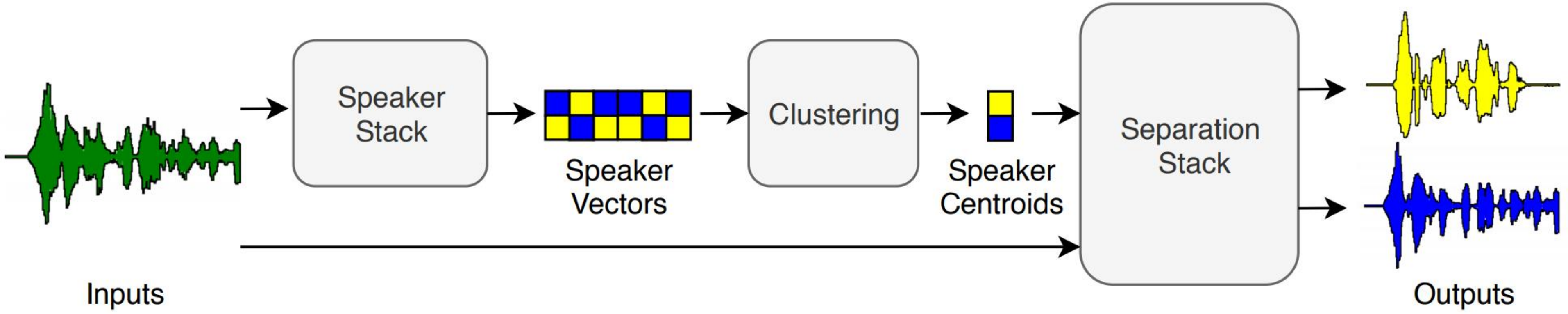


Figure 1. Wavesplit for 2-speaker separation. The speaker stack extracts speaker vectors at each timestep. The vectors are clustered and aggregated in speaker centroids. The separation stack ingests the centroids and the input signal to output two clean channels.

TasNet: variants on objective

MulCat DPRNN [8]: joint speaker identification and speaker separation

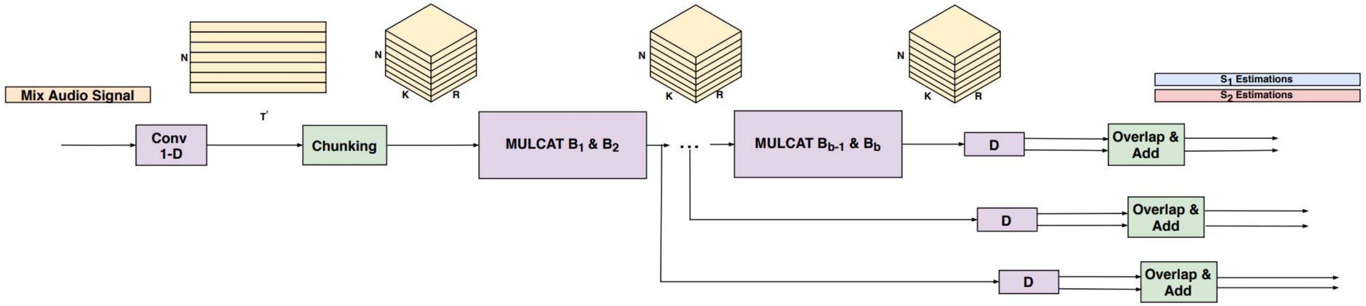


Figure 1. The architecture of our network. The audio is being convolved with a stack of 1D convolutions and reordered by cutting overlapping segments of length K in time, to obtain a 3D tensor. In our method, the RNN blocks are of the type of multiply and add. After each pair of blocks, we apply a convolution D to the copy of the activations, and obtain output channels by reordering the chunks and then using the overlap and add operator.

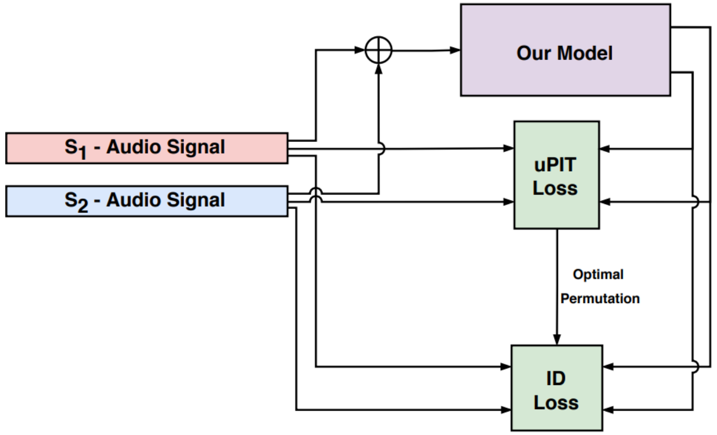


Figure 3. The training losses used in our method, shown for the case of $C = 2$ speakers. The mixed signal x combines the two input voices s_1 and s_2. Our model then separates to create two output channels $\hat{s}_1$ and $\hat{s}_2$. The permutation invariant SI-SNR loss computes the SI-SNR between the ground truth channels and the output channels, obtained at the channel permutation π that minimizes the loss. The identity loss is then applied to the matching channels, after they have been ordered by π.

TasNet: variants on task

Speech extraction: extract a target speaker from a mixture

- Get rid of the output dimension problem and the permutation problem

SpEx/SpEx+ [9-10]: extracting target speaker from a (trainable) speaker embedding

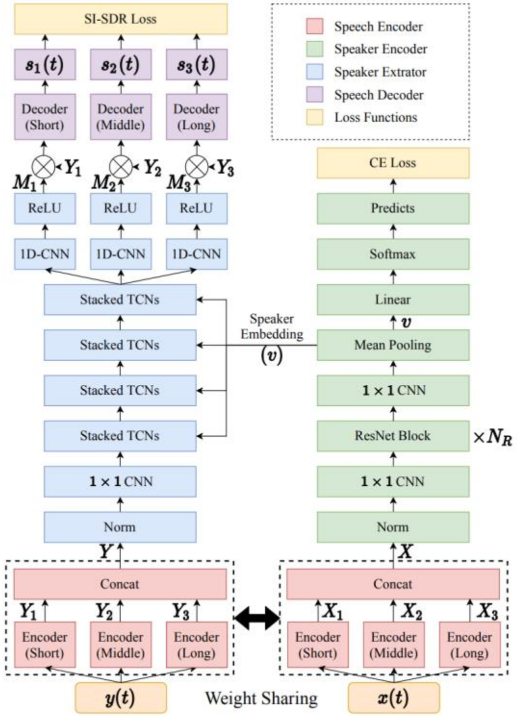


Figure 1: *The diagram of the proposed SpEx+ system, which consists of two twin speech encoders, speaker encoder, speaker extractor, and speech decoder. "Stacked TCNs" represents a stack of several TCN blocks (i.e. 8 TCN blocks in this work) with exponential dilation increment. The details of TCN block and ResNet block are shown in Fig. 2. The extracted signal* $s_1(t)$ *is chosen as the ultimate output at run-time inference.*

TasNet: variants on task

Music separation [11]: higher sample rate and source-specific model parameters

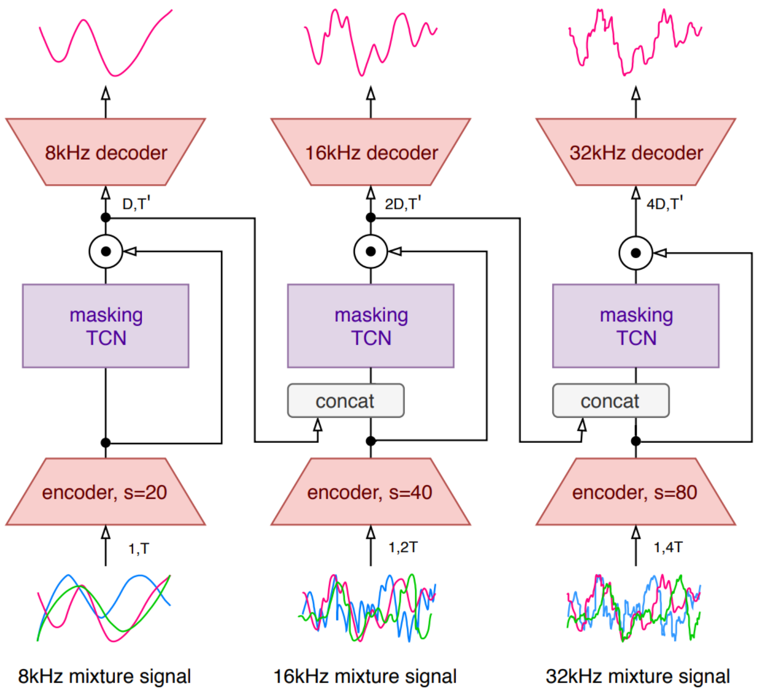


Fig. 2: Illustration of the multi-stage architecture. The resolution of the estimated signal is progressively enhanced by utilizing information from previous stages. The encoders increase the stride s to preserve the same time dimension T'.

TasNet: variants on task

Universal sound separation [12-13]: going beyond speech and music

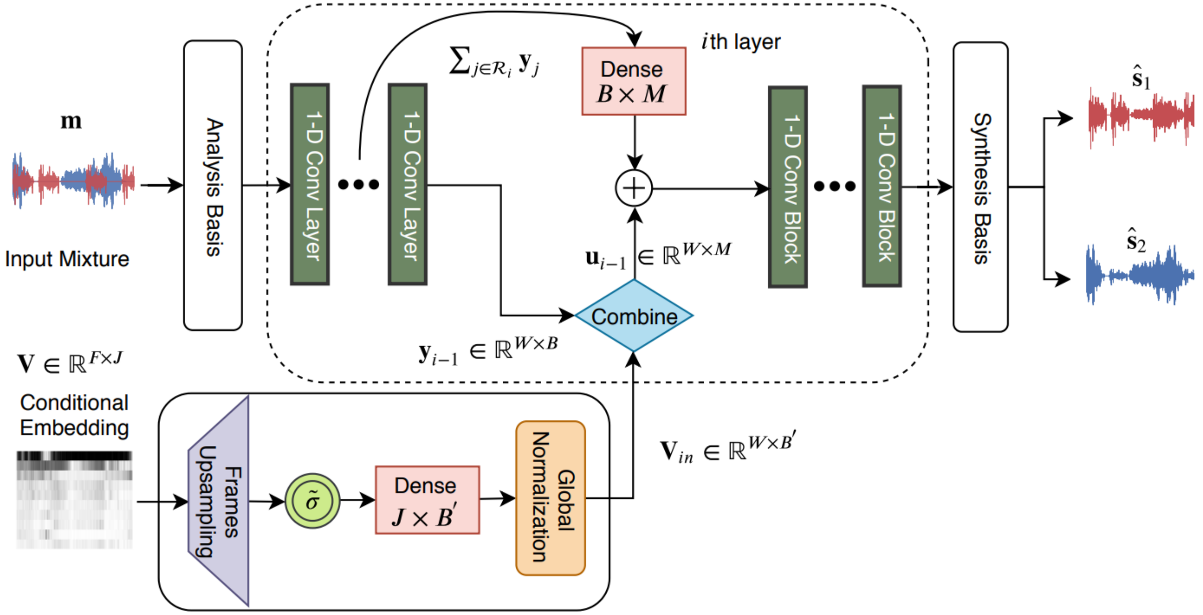


Fig. 2: Integrating conditional embedding information to the ith layer of a TDCN++ separation module.

TasNet: variants on task

Audio-visual separation [14]: exploring multi-model applications

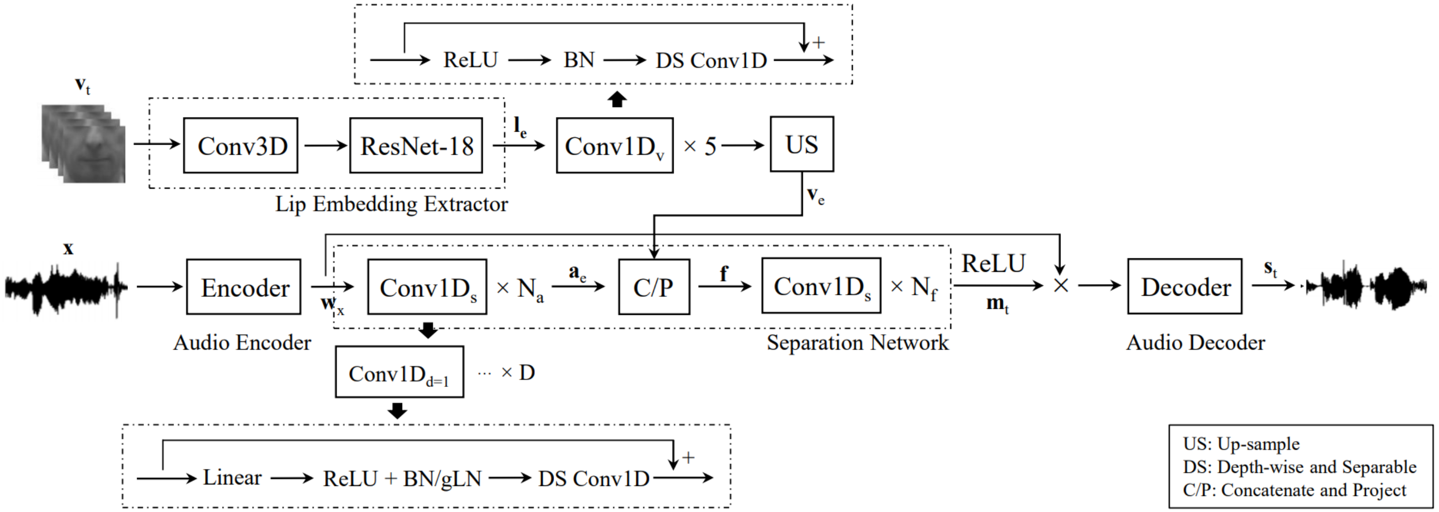


Fig. 1. Proposed time-domain audio-visual separation network. The lip region of each input image frame in $\mathbf{v}_t$ is cropped out and resized to 112 × 112 before feeding into the spatiotemporal convolutional block Conv3D [18]. Lip embedding $\mathbf{l}_e$, audio encoder output $\mathbf{w}_x$, video encoder output $\mathbf{v}_e$, audio encoded feature $\mathbf{a}_e$ and fused feature $\mathbf{f}$ are 256 dimensional sequences.

Single-channel systems: waveform-to-waveform

Directly mapping mixture waveform to target waveforms: “pure” end-to-end approach

- Regression-style estimation
- Mostly focus on the model design (typically 1D ConvNets)

Single-channel systems: waveform-to-waveform

Wave-U-Net [15]: applying U-net architecture to waveforms

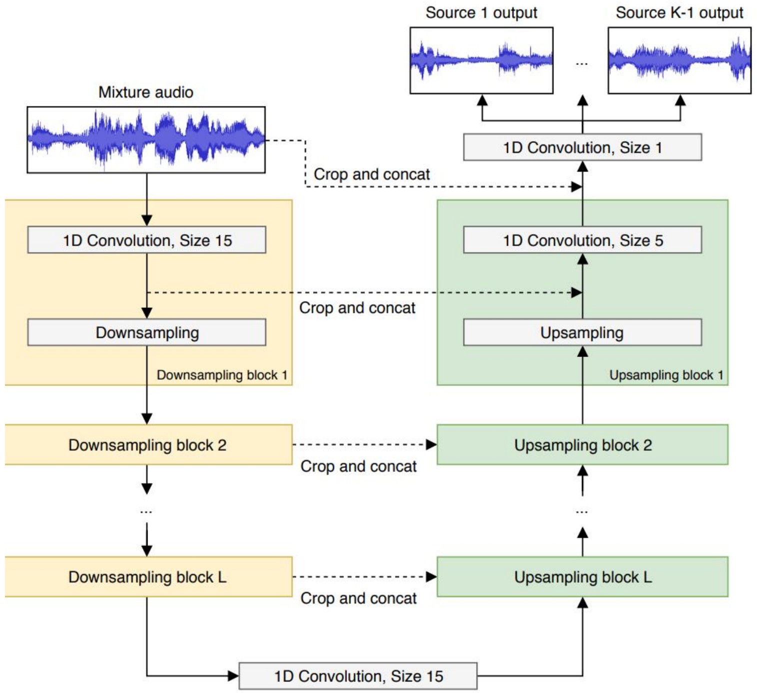


Figure 1. Our proposed Wave-U-Net with K sources and L layers. With our difference output layer, the K-th source prediction is the difference between the mixture and the sum of the other sources.

Single-channel systems: waveform-to-waveform

WaveNet-style model [16]: inspired by WaveNet

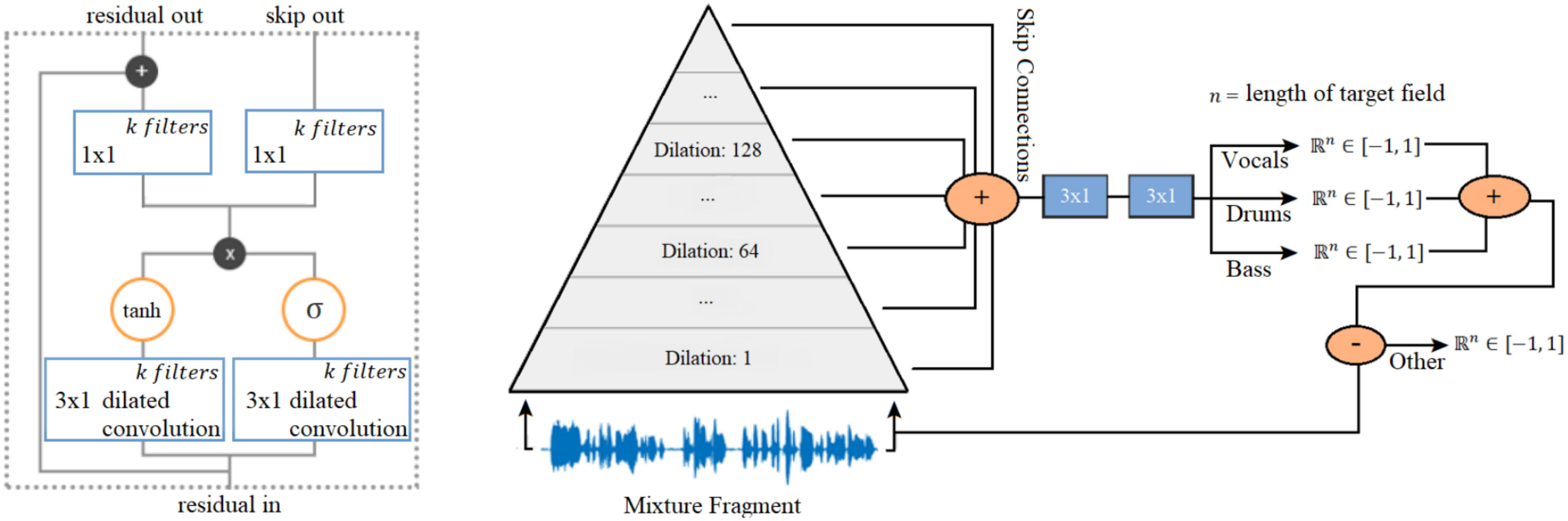


Figure 1: ***Left*** *– Residual layer.* ***Right*** *– Overview of the non-causal Wavenet we propose for multi-instrument source separation.*

Outline

End-to-end neural source separation: definition and progress

- Journey to the state-of-the-art

Single-channel systems

- Explicit replacement of short-time Fourier transform (STFT)
- Waveform-to-waveform mapping

Multi-channel systems

- Neural beamformers
- Single-channel extensions

Challenges and future directions

Multi-channel systems: overview

Multi-channel separation:

- Observation: $y_i = \sum_{j=1}^{C} x_i^j + n_i, i = 1, \dots, K \ (K > 1)$
- Target: x_1^j (or signals from certain directions)

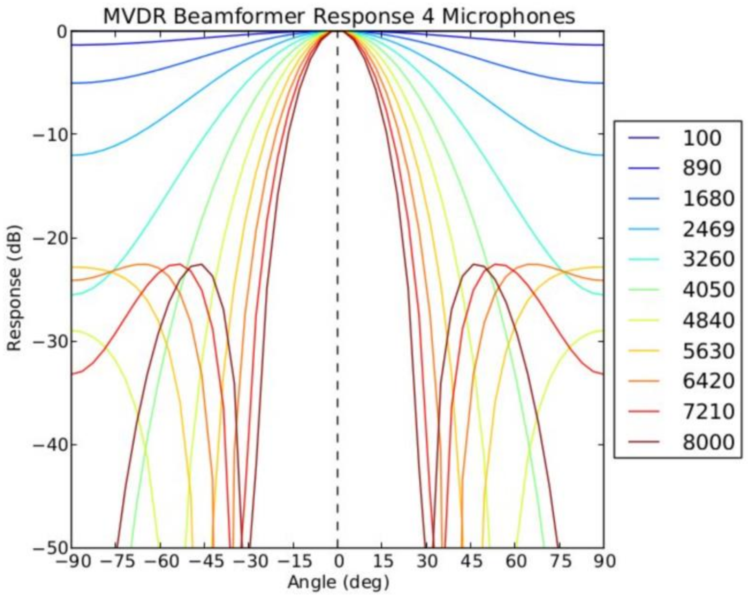
Conventional (mainstream) multi-channel method: **beamforming**

- Spatial filters that focus on signal coming from a certain direction

Filter calculation: estimate the target direction, then solve an optimization problem

- MMSE, MaxSNR, etc.
- Constraints may apply (e.g. distortionless response in the target direction)

Example beampattern of a minimum variance distortionless response (MVDR) beamformer:



Picture source: https://www.vocal.com/beamforming-2/minimum-variance-distortionless-response-mvdr-beamformer/

Multi-channel systems: neural beamformers

Neural beamformers: boost conventional beamformers with neural networks

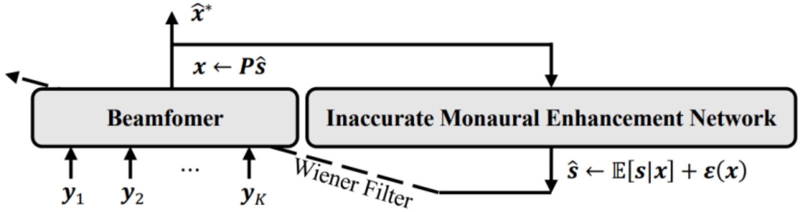
Three formulations:

- Mask-based beamformer: use single-channel T-F masks to estimate beamforming filters (frequency domain)
- Output-based beamformer: use single-channel separation outputs to estimate beamforming filters
- DNN beamformer: use neural networks to directly estimate beamforming filters

End-to-end (time-domain) beamformers: output-based and DNN

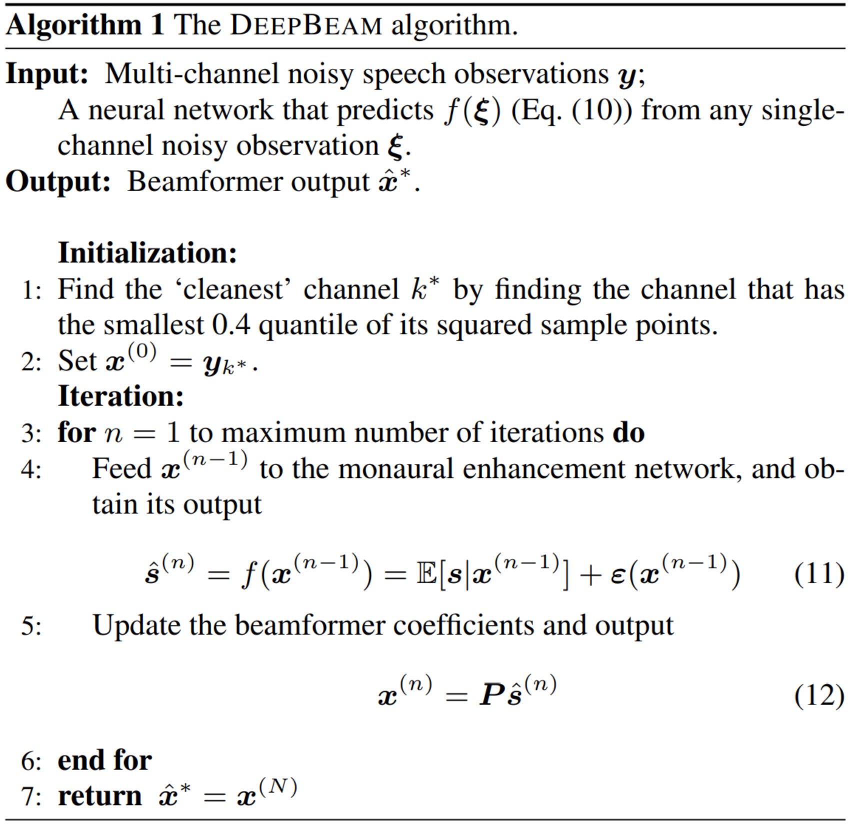
Multi-channel systems: output-based beamformers

DeepBeam [17]: Single-channel enhancement + time-domain multi-channel Wiener filtering (MWF):



$\hat{x}^*$
$x \leftarrow P\hat{s}$
Beamfomer
Inaccurate Monaural Enhancement Network
y_1
y_2
...
y_K
Wiener Filter
$\hat{s} \leftarrow \mathbb{E}[s|x] + \varepsilon(x)$

Fig. 1: DeepBeam framework.



Algorithm 1 The DeepBeam algorithm.

```
Input: Multi-channel noisy speech observations $\boldsymbol{y}$;
    A neural network that predicts $f(\boldsymbol{\xi})$ (Eq. (10)) from any single-channel noisy observation $\boldsymbol{\xi}$.
Output: Beamformer output $\hat{\boldsymbol{x}}^*$.

Initialization:
Find the 'cleanest' channel $k^*$ by finding the channel that has the smallest 0.4 quantile of its squared sample points.
Set $\boldsymbol{x}^{(0)} = \boldsymbol{y}_{k^*}$.
Iteration:
for $n = 1$ to maximum number of iterations do
    Feed $\boldsymbol{x}^{(n-1)}$ to the monaural enhancement network, and obtain its output
        $\hat{\boldsymbol{s}}^{(n)} = f(\boldsymbol{x}^{(n-1)}) = \mathbb{E}[\boldsymbol{s}|\boldsymbol{x}^{(n-1)}] + \boldsymbol{\varepsilon}(\boldsymbol{x}^{(n-1)})$    (11)
    Update the beamformer coefficients and output
        $\boldsymbol{x}^{(n)} = \boldsymbol{P}\hat{\boldsymbol{s}}^{(n)}$    (12)
end for
return $\hat{\boldsymbol{x}}^* = \boldsymbol{x}^{(N)}$
```

Multi-channel systems: output-based beamformers

Beam-TasNet [18]: single-channel TasNet outputs + frequency-domain MVDR beamforming

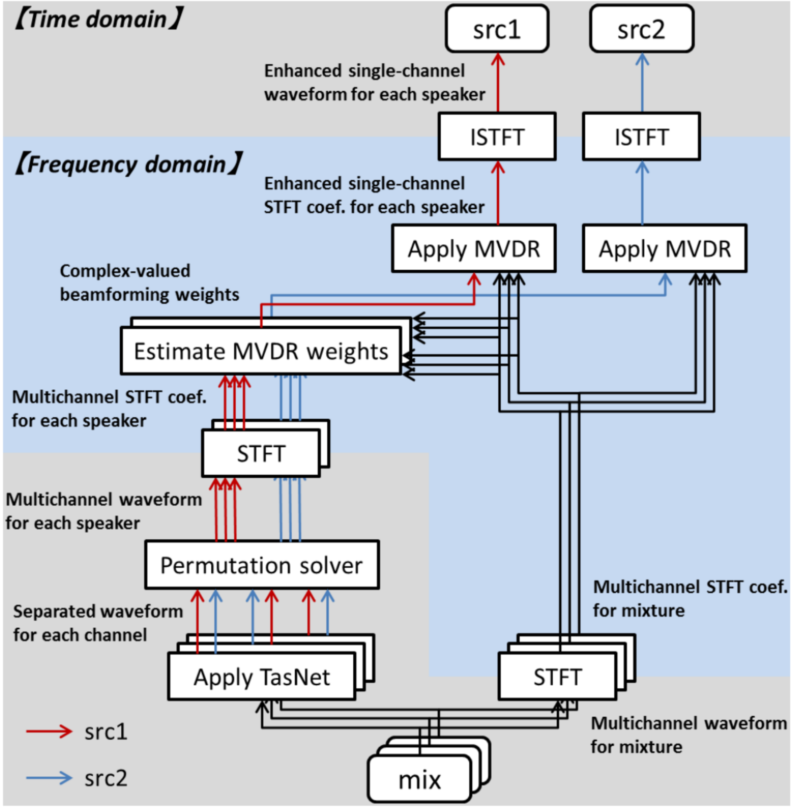

【Time domain】
src1
src2
Enhanced single-channel waveform for each speaker
ISTFT
ISTFT
【Frequency domain】
Enhanced single-channel STFT coef. for each speaker
Apply MVDR
Apply MVDR
Complex-valued beamforming weights
Estimate MVDR weights
Multichannel STFT coef. for each speaker
STFT
Multichannel waveform for each speaker
Permutation solver
Separated waveform for each channel
Multichannel STFT coef. for mixture
Apply TasNet
STFT
Multichannel waveform for mixture
src1
src2
mix

Fig. 1. Overall separation procedures in Beam-TasNet

Multi-channel systems: DNN beamformers

Neural network adaptive beamforming (NAB) [19]:
time-domain beamforming with joint ASR training

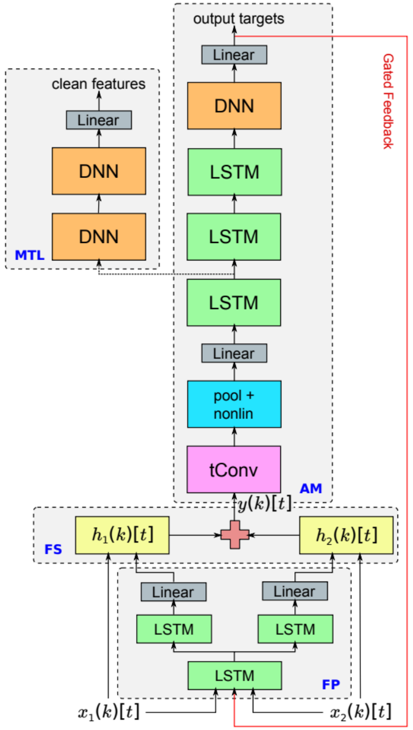
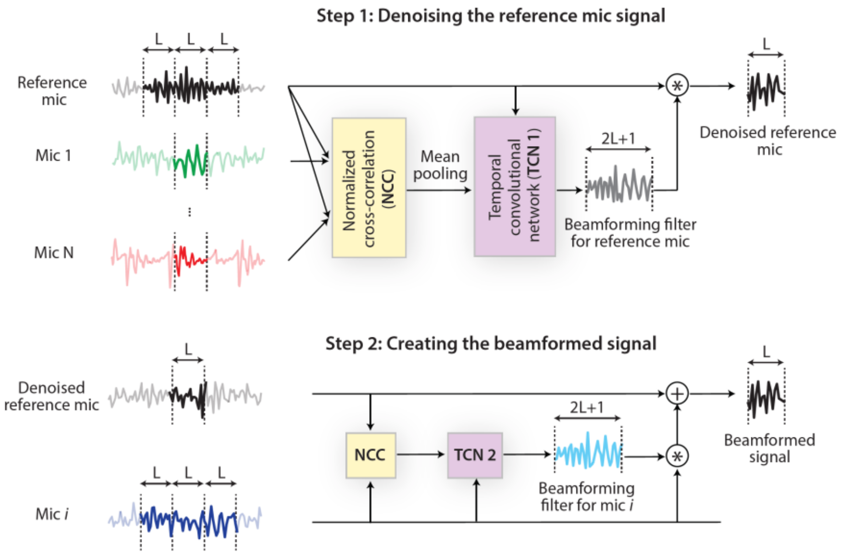
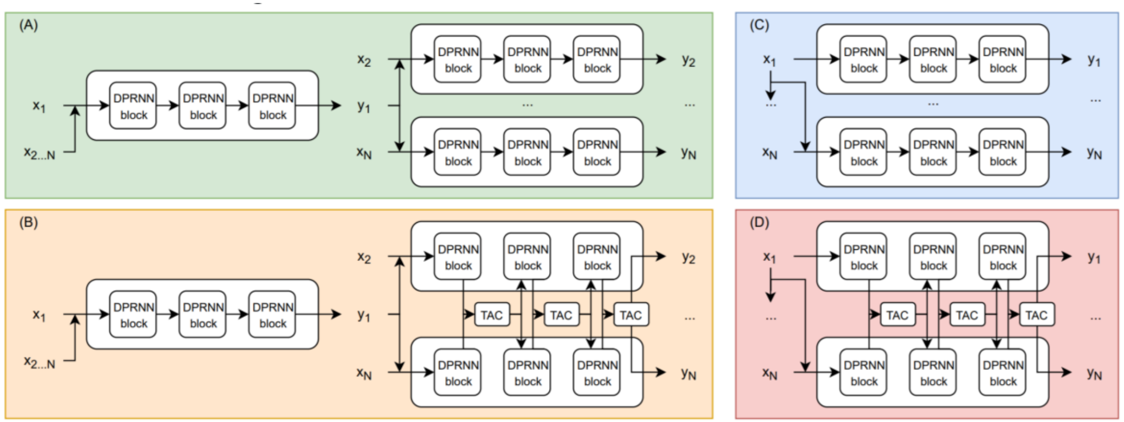


Figure 1: *Neural network adaptive beamforming (NAB) model architecture. It consists of filter prediction (FP), filter-and-sum (FS) beamforming, acoustic modeling (AM) and multitask learning (MTL) blocks. Only two channels are shown for simplicity.*

Multi-channel systems: DNN beamformers

Filter-and-sum network (FaSNet) [20-21]: directly estimate time-domain beamforming filters


Step 1: Denoising the reference mic signal
L
L
L
Reference mic
Mic 1
Mic N
Normalized cross-correlation (NCC)
Mean pooling
Temporal convolutional network (TCN 1)
2L+1
Beamforming filter for reference mic
L
Denoised reference mic
Step 2: Creating the beamformed signal
Denoised reference mic
Mic i
NCC
TCN 2
2L+1
Beamforming filter for mic i
Beamformed signal



Multi-channel systems: single-channel extensions

Extend single-channel models to receive:

- Single-channel input with cross-channel features
- Multi-channel inputs

It can be either “masking-based” (TasNet-style) or direct mapping

Multi-channel systems: single-channel extensions

Multi-channel TasNet [22-23]: append (conventional or learned) cross-channel features to single-channel TasNet

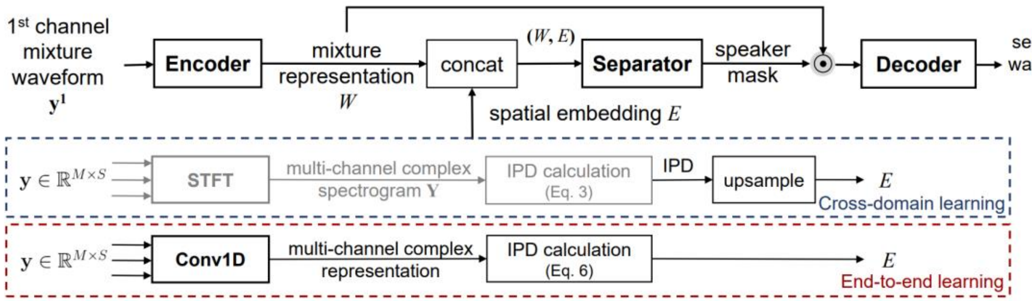

1st channel mixture waveform $\mathbf{y}^1$
Encoder
mixture representation W
concat
(W, E)
Separator
speaker mask
Decoder
separated waveforms $\hat{\mathbf{s}}$
spatial embedding E
$\mathbf{y} \in \mathbb{R}^{M \times S}$
STFT
multi-channel complex spectrogram $\mathbf{Y}$
IPD calculation (Eq. 3)
IPD
upsample
E
Cross-domain learning
$\mathbf{y} \in \mathbb{R}^{M \times S}$
Conv1D
multi-channel complex representation
IPD calculation (Eq. 6)
E
End-to-end learning

Figure 2: *The block diagram of our proposed multi-channel speech separation networks. The dashed boxes denotes the method described in Section 3.1 and Section 3.2, respectively. The blocks marked in grey indicate they are not involved in network training.*

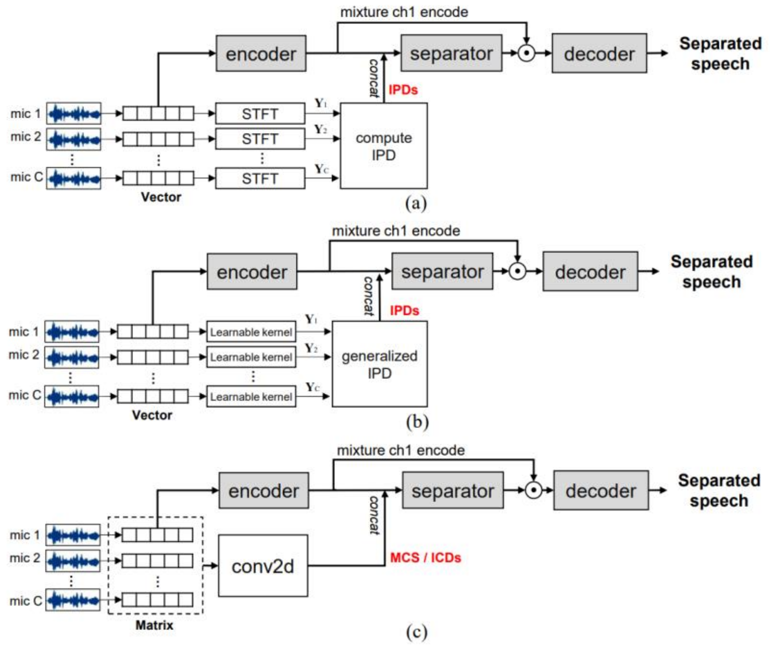


Fig. 1. The diagram of MCSS model (a) incorporating with IPDs computed by standard STFT. (b) incorporating with generalized IPDs computed by learnable STFT kernel [8]. (c) incorporating with our proposed MCS and ICDs computed by a conv2d layer.

Multi-channel systems: single-channel extensions

Multi-channel Wave-U-Net [15]: accept multiple input channels

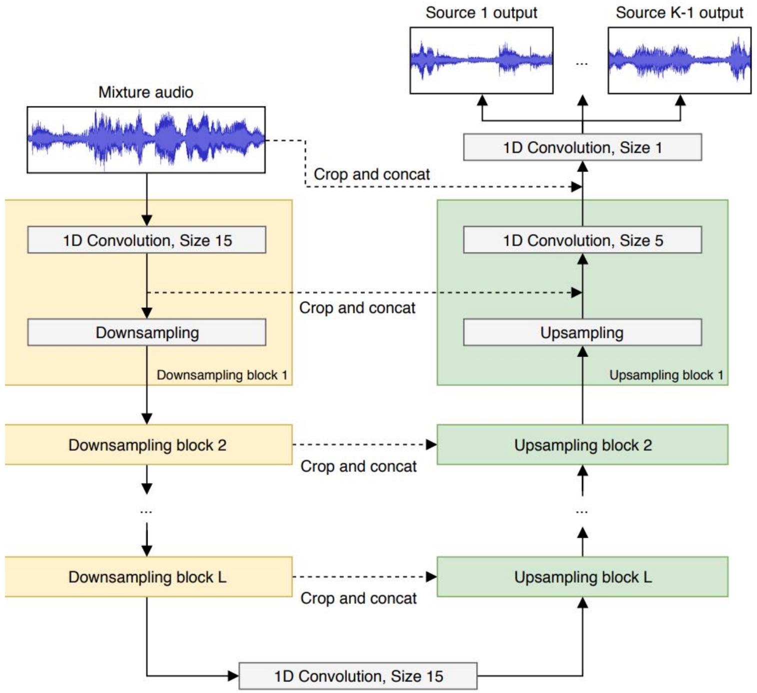


Figure 1. Our proposed Wave-U-Net with K sources and L layers. With our difference output layer, the K-th source prediction is the difference between the mixture and the sum of the other sources.

Outline

End-to-end neural source separation: definition and progress

- Journey to the state-of-the-art

Single-channel systems

- Explicit replacement of short-time Fourier transform (STFT)
- Waveform-to-waveform mapping

Multi-channel systems

- Neural beamformers
- Single-channel extensions

Challenges and future directions

Challenges and future directions

Source counting and output permutation

- How many sources are there in the mixture?
  - Multi-talker vocal activity detection (VAD) [24], speaker diarization [25], etc.
  - Proper training objectives [26-27]

Continuous separation on long-span audio

- Session-level (1~10 min, or 1 hour) separation instead of utterance-level separation
  - LibriCSS dataset for meeting-wise processing [28]
- A topic of JHU JSALT 2020 workshop (ongoing) (https://www.clsp.jhu.edu/speech-recognition-and-diarization-for-unsegmented-multi-talker-recordings-with-speaker-overlaps/)
  - Dynamic source counting, separation, and diarization
  - Long-sequence modeling

Challenges and future directions

Universal separation: the ultimate goal [12-13]

Frontend-backend fusion: end-to-end separation and recognition [29-30]

Real-world environments: noise and reverberation [31], domain mismatch

References

[1] Le Roux, Jonathan, et al. "SDR–half-baked or well done?." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[2] Luo, Yi, and Nima Mesgarani. "TasNet: time-domain audio separation network for real-time, single-channel speech separation." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[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 27.8 (2019): 1256-1266.

[4] Luo, Yi, Zhuo Chen, and Takuya Yoshioka. "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[5] Tzinis, Efthymios, et al. "Two-Step Sound Source Separation: Training On Learned Latent Targets." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[6] Ditter, David, and Timo Gerkmann. "A multi-phase gammatone filterbank for speech separation via tasnet." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[7] Zeghidour, Neil, and David Grangier. "Wavesplit: End-to-end speech separation by speaker clustering." arXiv preprint arXiv:2002.08933 (2020).

[8] Nachmani, Eliya, Yossi Adi, and Lior Wolf. "Voice Separation with an Unknown Number of Multiple Speakers." arXiv preprint arXiv:2003.01531 (2020).

References

[9] Xu, Chenglin, et al. "SpEx: Multi-Scale Time Domain Speaker Extraction Network." arXiv preprint arXiv:2004.08326 (2020).

[10] Ge, Meng, et al. "SpEx+: A Complete Time Domain Speaker Extraction Network." arXiv preprint arXiv:2005.04686 (2020).

[11] Samuel, David, Aditya Ganeshan, and Jason Naradowsky. "Meta-learning Extractors for Music Source Separation." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[12] Kavalerov, Ilya, et al. "Universal sound separation." 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2019.

[13] Tzinis, Efthymios, et al. "Improving universal sound separation using sound classification." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[14] Wu, Jian, et al. "Time domain audio visual speech separation." arXiv preprint arXiv:1904.03760 (2019).

[15] Stoller, Daniel, Sebastian Ewert, and Simon Dixon. "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation." arXiv preprint arXiv:1806.03185 (2018).

[16] Lluís, Francesc, Jordi Pons, and Xavier Serra. "End-to-end music source separation: is it possible in the waveform domain?." arXiv preprint arXiv:1810.12187 (2018).

References

[17] Qian, Kaizhi, et al. "Deep learning based speech beamforming." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[18] Ochiai, Tsubasa, et al. "Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[19] Li, Bo, et al. "Neural network adaptive beamforming for robust multichannel speech recognition." (2016).

[20] Luo, Yi, et al. "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing." 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019.

[21] Luo, Yi, et al. "End-to-end microphone permutation and number invariant multi-channel speech separation." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[22] Gu, Rongzhi, et al. "End-to-end multi-channel speech separation." arXiv preprint arXiv:1905.06286 (2019).

[23] Gu, Rongzhi, et al. "Enhancing End-to-End Multi-Channel Speech Separation Via Spatial Feature Learning." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[24] Medennikov, Ivan, et al. "Target-Speaker Voice Activity Detection: a Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario." arXiv preprint arXiv:2005.07272 (2020).

References

[25] Horiguchi, Shota, et al. "End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors." arXiv preprint arXiv:2005.09921 (2020).

[26] Takahashi, Naoya, et al. "Recursive speech separation for unknown number of speakers." arXiv preprint arXiv:1904.03065 (2019).

[27] Luo, Yi, and Nima Mesgarani. "Separating varying numbers of sources with auxiliary autoencoding loss." arXiv preprint arXiv:2003.12326 (2020).

[28] Chen, Zhuo, et al. "Continuous speech separation: Dataset and analysis." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[29] von Neumann, Thilo, et al. "End-to-end training of time domain audio separation and recognition." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[30] von Neumann, Thilo, et al. "Multi-talker ASR for an unknown number of sources: Joint training of source counting, separation and ASR." arXiv preprint arXiv:2006.02786 (2020).

[31] Maciejewski, Matthew, et al. "WHAMR!: Noisy and reverberant single-channel speech separation." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

Q&A

1. 单通道与多通道语音分离的各自应用场景有什么不同，多通道有什么具体优势吗

2.多通道的评估指标该怎么计算，因为源文件和混合音频文件有多个通道，分离结果该怎么像单通道一样进行比对呢

3.目前的语音分离还有哪些不足还需完善才能落地商用呢，现在有online（实时的）语音分离的研究吗

4.这些方法可以用来单通道语音降噪嘛，如果可以的话在实时性上和rnnoise算法相比怎么样，模型的大小和速度和效果上如何？

5提问：Audio-Visual Speech Separation的研究进展如何？未来的研究热点集中在哪儿

6.基于深度学习的降噪，什么样的代价函数效果最好，像基于irm的mse以及luoyi老师用的si-snr效果差异是否很大

Q&A

7.问题：相比于频域，是不是基于时域的分离上限一定会更高?

8.近期的研究大多偏向时域分离，那时域分离效果为什么优于频域效果呢？以后可能会向什么方向发展呢

9.使用wsj数据训练的分离模型，在wsj测试集上性能很好。但如果跨信道，比如在安卓手机录音上测试，性能会出现较大下降。请问怎么解决跨信道的问题?

10.我这边复现的两个语音分离模型在一个数据集上训练并测试效果良好，但是换一个数据集去测试后，效果大幅下降，是我工作出了问题，还是这是普遍存在的问题？如果普遍存在，可能的原因和解决方案有哪些?

11.source separation和speaker-diarization在技术实现上有什么联系吗，我在做speaker-diarization任务，能借鉴source separation的什么思路?

Q&A

12. 对于刚开始接触语音分离的研究生，应该如何循序渐进地学习语音分离呢？有推荐的书籍或者课程来学习语音分离嘛

13.当输入音频存在混响的时候，对于Si-SNR loss 是否需要修正？

14.说话人分离相比于普通语音分离(噪声)的难点以及现有的最好说话人分离系统是什么

15.目前语音分离技术的进展，有了解到工业或者哪个产品上落地吗

16.在多通道分离中，IPD常常作为辅助特征来引导语音分离，请问它的具体物理意义是什么呢？

17.现在时域上直接对语音处理效果很好，是不是代表频域上不行了呢?

18.客服电话包含用户和客服的语音分离，一般采用什么方法？如何避免角色识别错误和端点检测不准确问题？

Q&A

19.盲源分离和降噪有哪些比较新的idea和应用？

20.多通道语音分离中有哪些合适的数据集

21.请问语音分离技术可以用在啸叫抑制和回声消除吗

22.单通道语音分离由于没有IPD/ISD等空间信息，只能做谱分析，那么说话人的声纹特征是否是主要/重要的分离依据？训练集中不同说话人样本太少是否是跨数据集表现大幅下降的主要原因？如果把speaker-diarization任务中的一些pre-train模块放到分离网络前辅助encode是否会有提升？

23.干扰人声，混响，和噪声，远场，能不能一起弄了？不同重叠率的泛化性能怎么提升？

24. 多种声音混叠时，要很好地分离声源，有没有混叠源数的上限？我看现在都是2个人、然后无重叠的分离较多