



Speech Denoising

Héctor Martel 何可拓 Samuel Pegg 茄子 Toghrul Abbasli 吐谷鲁

{hkt20, peggsr10, abbasli10}@mails.tsinghua.edu.cn



Problem definition

- Real-world speech contains noise
- Objective: Remove the background noises to **clean the speech signal**

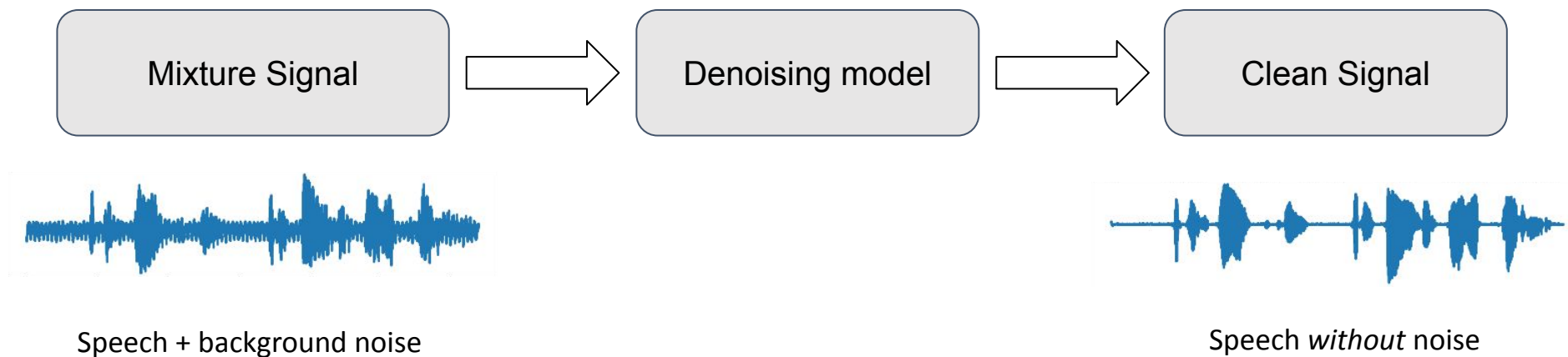




Problem definition

- Real-world speech contains noise
- Objective: Remove the background noises to **clean the speech signal**
- Assumption: **noise is additive**

Enhancement approach

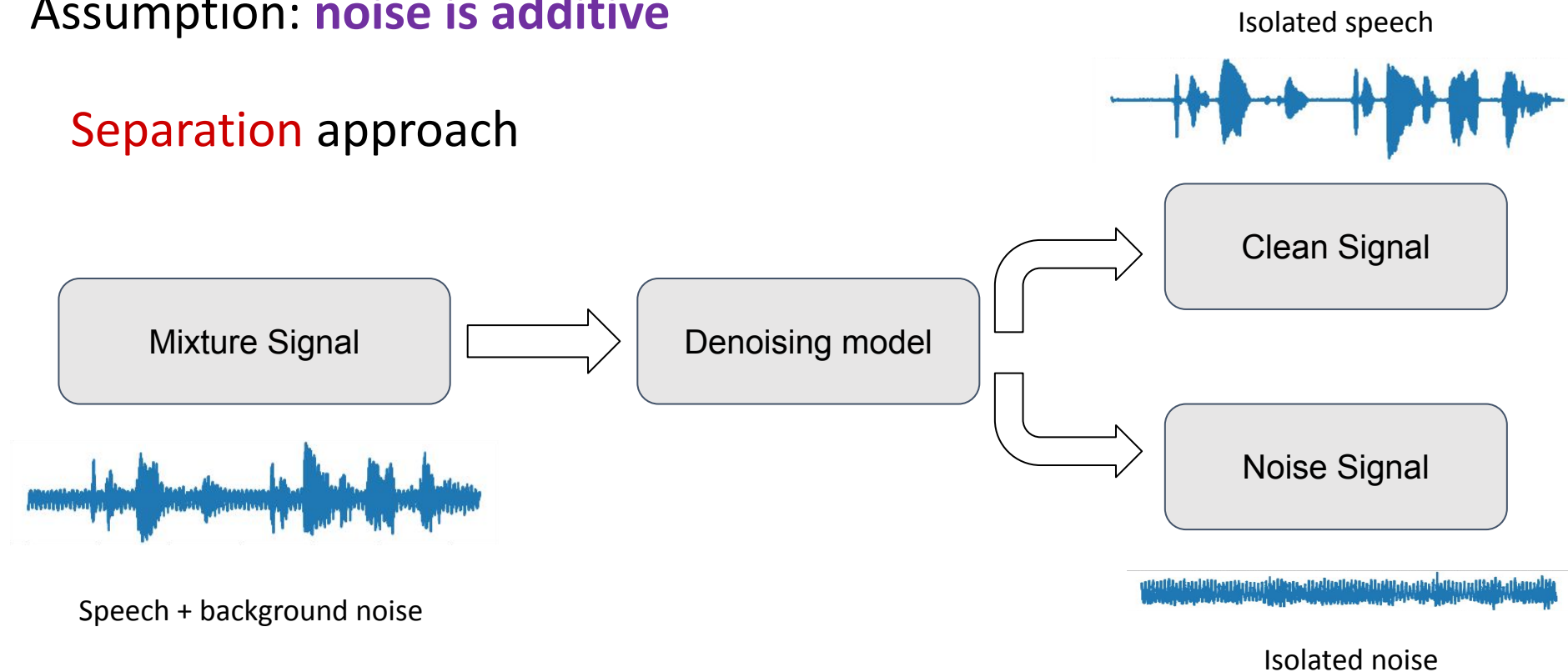




Problem definition

- Real-world speech contains noise
- Objective: Remove the background noises to **clean the speech signal**
- Assumption: **noise is additive**

Separation approach





Data generation

LibriSpeech (train-clean-100)



- Designed for Speech Recognition
- 100h of speech recordings
- Audiobooks in English
- 251 speakers (126 male, 125 female)
- 25 min/speaker



UrbanSound8K



- 8,732 labeled environmental sounds
- 4 seconds approx.
- 10 noise classes:

Air conditioner	Car horn
Children playing	Dog bark
Drilling	Engine idling
Gun shot	Jackhammer
Siren	Street music



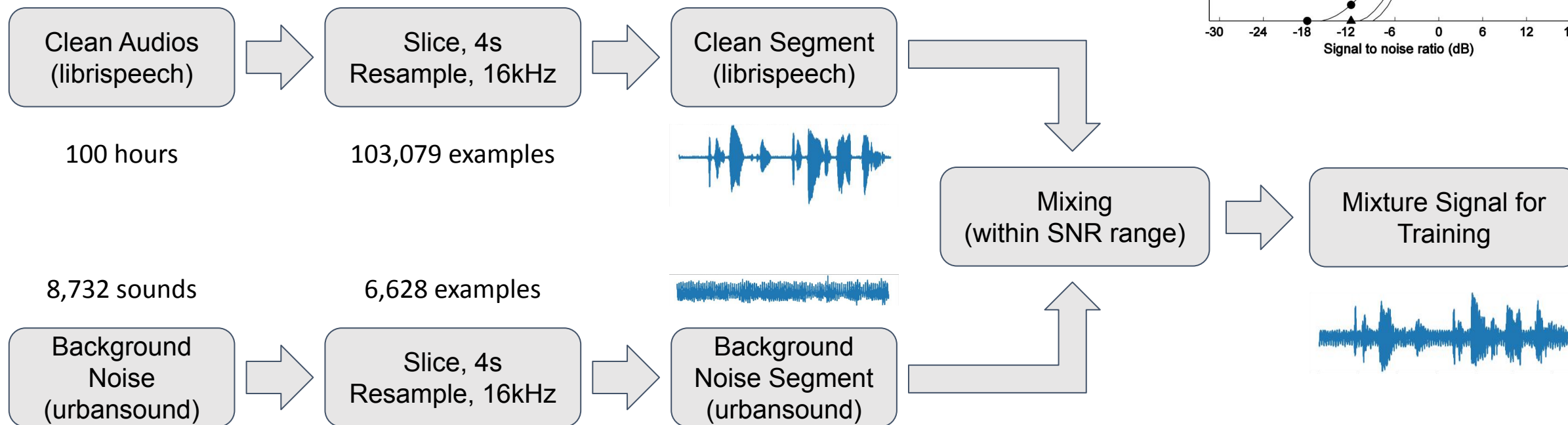
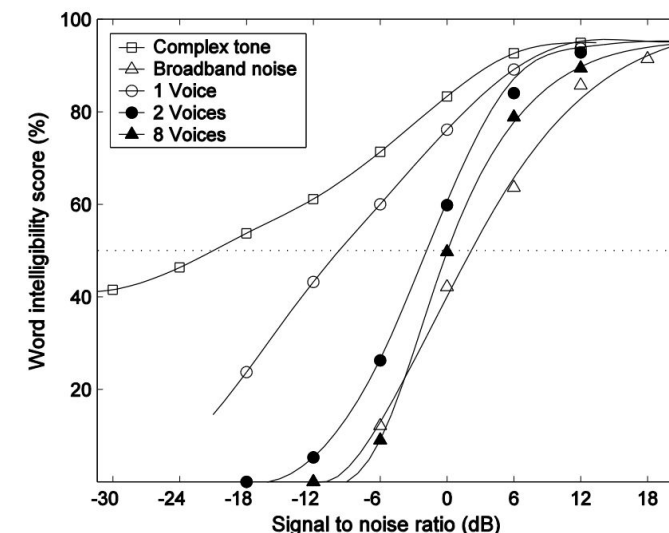
Data generation

Supervised Speech Separation Based on Deep Learning: An Overview, Wang and Chen (2017)

LibriSpeech (train-clean-100)



UrbanSound8K





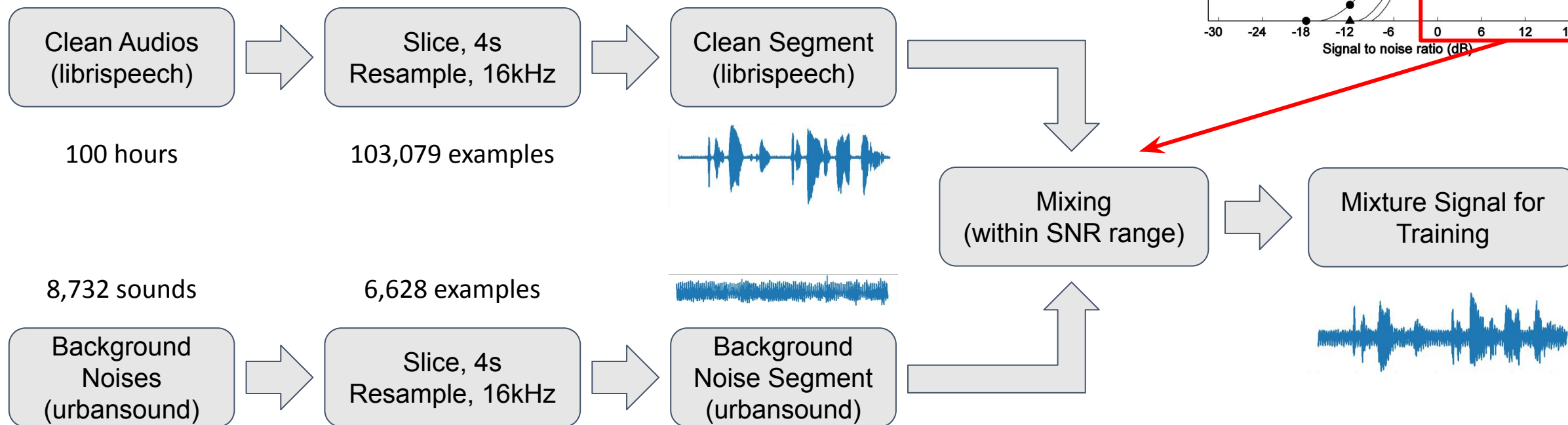
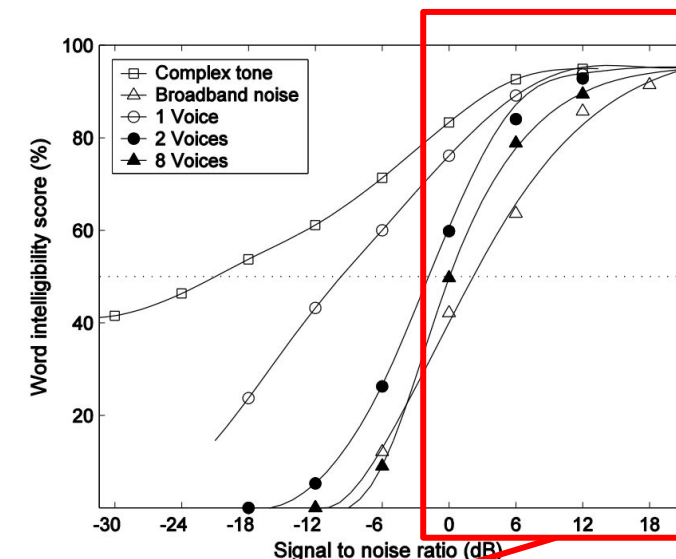
Data generation

Supervised Speech Separation Based on Deep Learning: An Overview, Wang and Chen (2017)

LibriSpeech (train-clean-100)

OpenSLR

UrbanSound8K



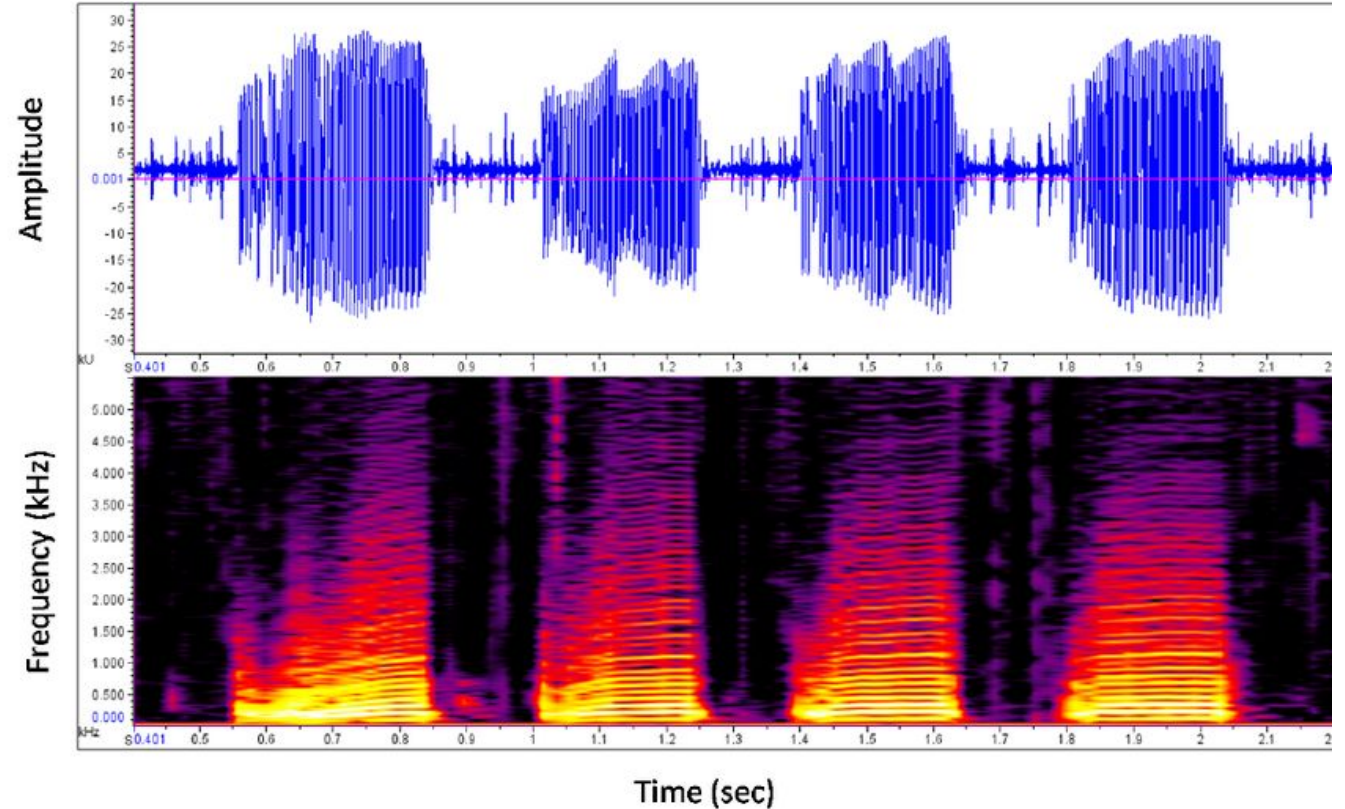


Time vs Frequency Domain

We can work directly with the audio signal, or convert the audio into a Frequency representation (2D).

Time Domain

Frequency Domain

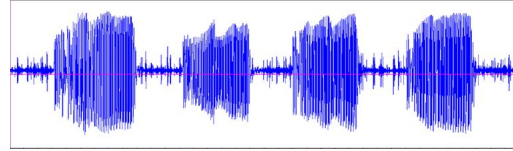




Time vs Frequency Domain

We can work directly with the audio signal, or convert the audio into a Frequency representation (2D).

Time Domain



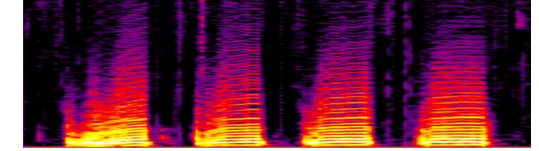
Pros

- Less preprocessing
- Avoid reconstruction errors

Cons

- Very long sequence of samples

Frequency Domain



Pros

- Converts problem to 2D: image processing techniques apply

Cons

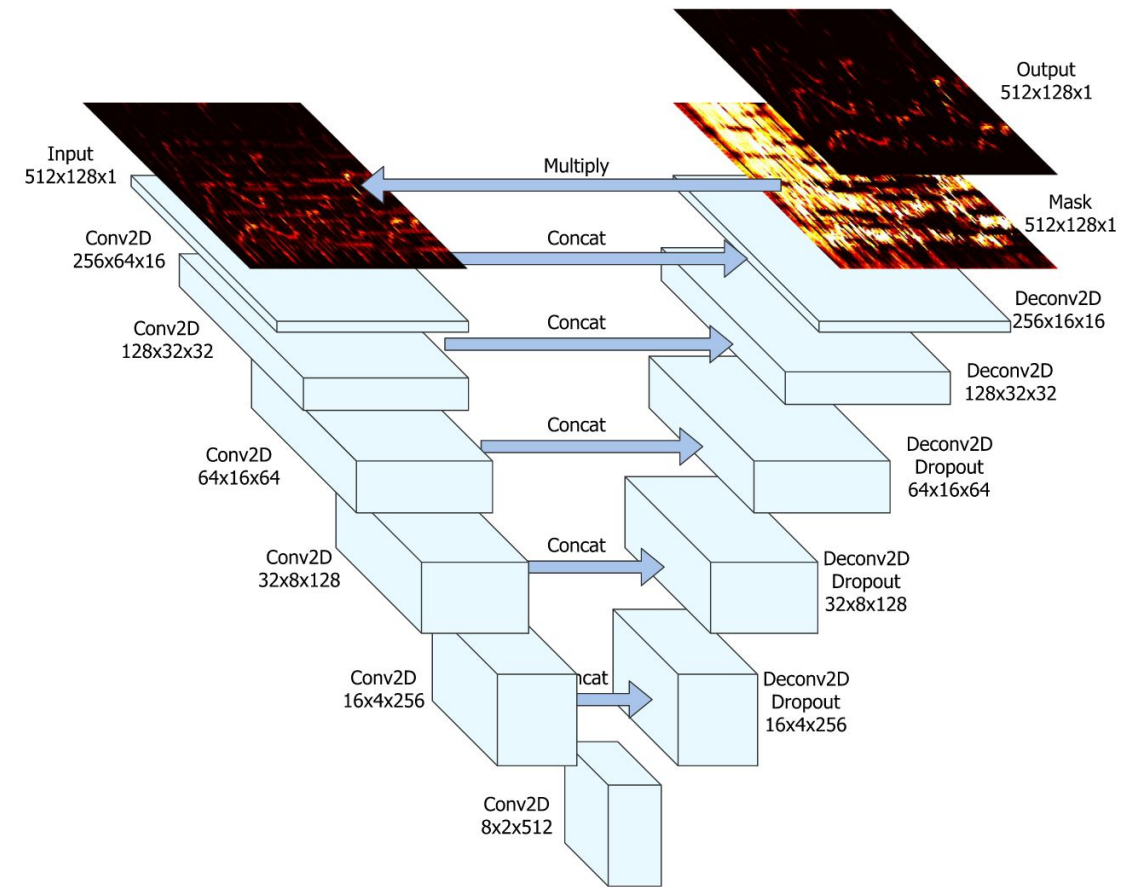
- Conversion can produce artifacts or loss of information
- Additional preprocessing time



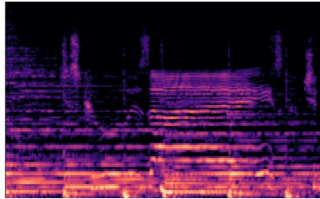
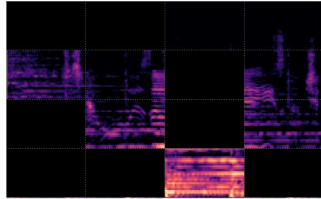
Methods: UNet

- CNN in Frequency domain
- Hyperparameters: depth, scalefactor between layers, dropout, *apply masks*
- Double convolution layers at each level
- Residual connections: Addition instead of concat as the aggregation for the layers
- Apply masks: either compute the output image directly, or calculate the **probability** of each pixel going to each class

U-Net: Convolutional Networks for Biomedical Image Segmentation



Mask			
0	0	1	1
1	1	0	1
0	1	0	1
0	0	1	0

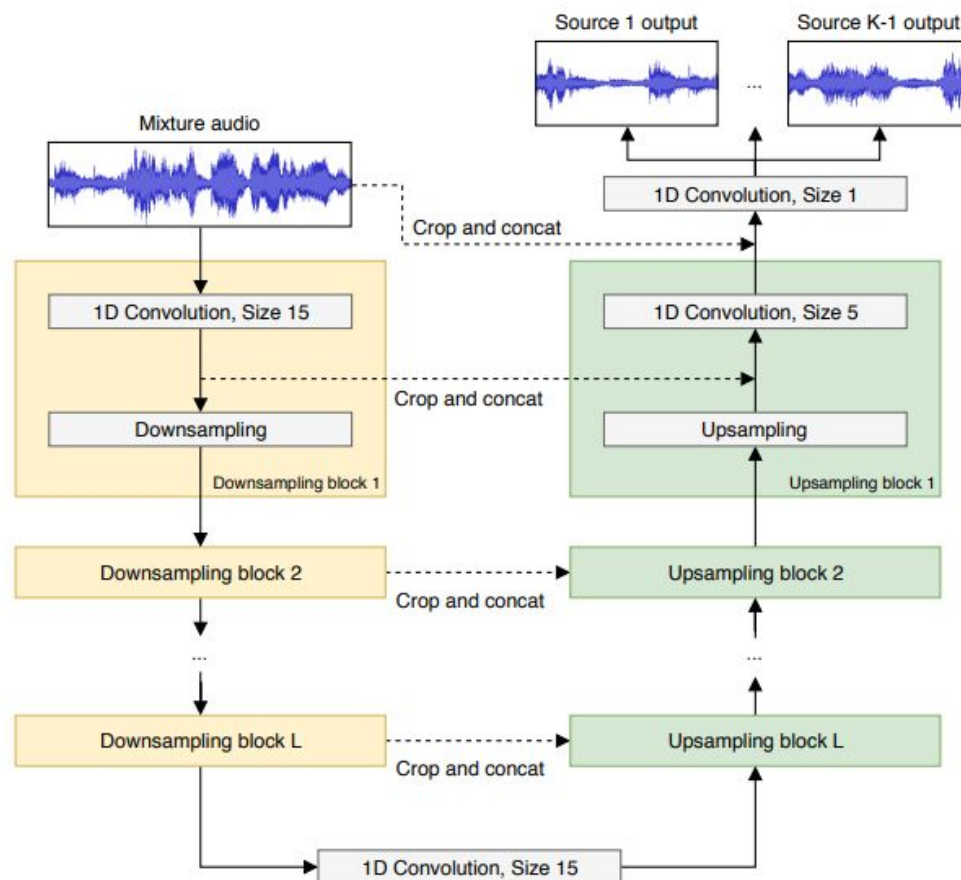
 \times  $=$ 



Methods: UNetDNP

- CNN in Time domain
- Adaptation of **WavUNet**: an implementation of UNet that replaces the 2D Convolutions with 1D Convolutions, and works directly on the time domain audio signal
- Original UNetDNP implementation is unsupervised, we adapted to the supervised setting
- **Direct estimation** of waveforms. We tried masking here too but it ended up introducing more artifacts

Speech Denoising by Accumulating
Per-Frequency Modeling Fluctuations

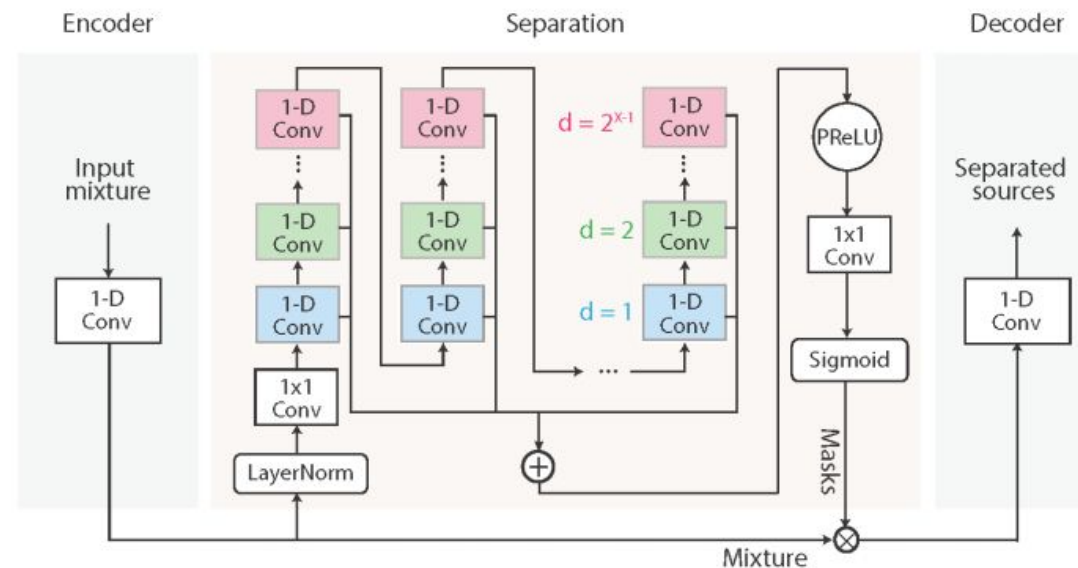
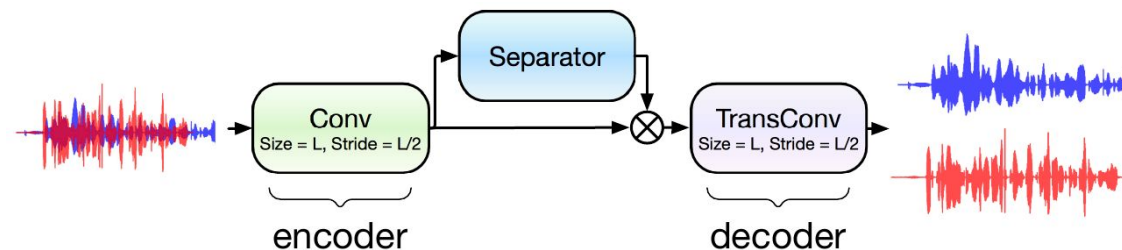




Methods: ConvTasNet

- CNN in Time domain
- Internally it uses an encoder-decoder structure to **learn a representation** of the audio signal in **base signals**
- It learns *“its own version of STFT and iSTFT”*
- Passes signal through several sets of convolutional layers. Output of each convolution fed into the next layer, and also saved separately. The saved outputs are then aggregated, combined with inputs, and sent to decoder
- Decoder returns the separated sources

Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation





Methods: TransUNet

- Transformer architecture in Frequency domain
- Structure similar to UNet. The “middle” section implements a feature extraction CNN + transformer block
- Transformer uses a MSA and MLP with skip-connections

TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation

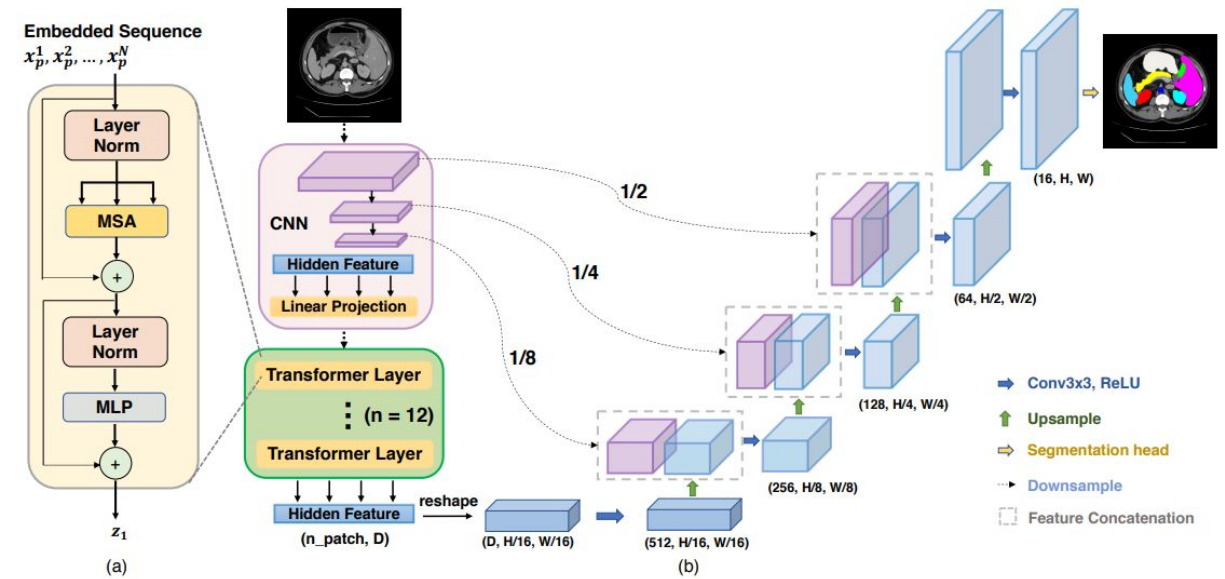


Fig. 1: Overview of the framework. (a) schematic of the Transformer layer; (b) architecture of the proposed TransUNet.



Evaluation metrics: Speech quality

- **Perceptual Evaluation of Speech Quality(PESQ)**
 - Correlated with human listening perception
 - Used to be an evaluation metric for phone networks and speech codecs
 - Aggregation of disturbance in frequency and time
- **Short-Time Objective Intelligibility (STOI)**
 - Intelligibility measure for speech data: “how easy to understand”
 - Local normalization for Time and Frequency units
 - Linear correlation coefficient between clean and processed Time and Frequency units



Results: objective evaluation

For all metrics, higher is better.

		Separation metrics			Speech quality metrics		
		SI-SDR (dB)	SI-SIR (dB)	SI-SAR (dB)	PESQ-wb	PESQ-nb	STOI
UNet	Frequency	-26.99	18.11	-26.98	1.29	1.74	0.75
UNetDNP	Time	7.61	16.90	8.55	1.56	2.07	0.82
ConvTasNet	Time	10.69	22.19	11.30	1.83	2.38	0.85
TransUNet **	Frequency	-27.83	12.24	-27.81	1.42	1.90	0.75

Table 1. Objective evaluation results for all models.
Values represent the mean across all evaluation examples.



Results: objective evaluation

For all metrics, higher is better.

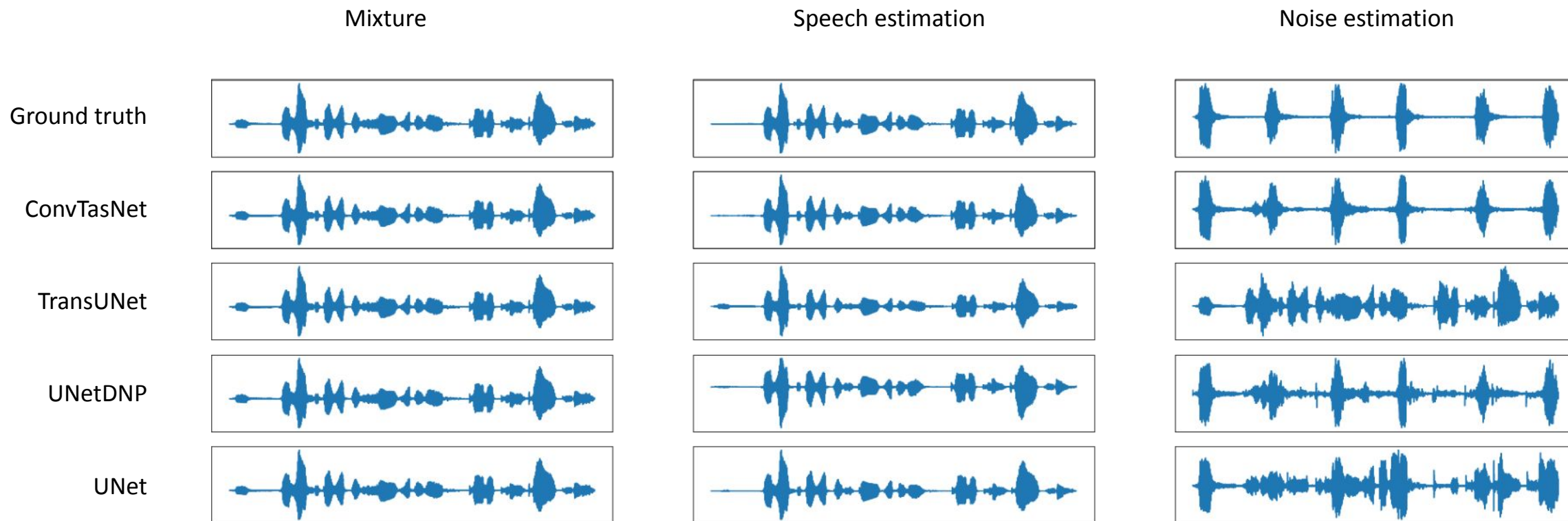
		Separation metrics			Speech quality metrics		
		SI-SDR (dB)	SI-SIR (dB)	SI-SAR (dB)	PESQ-wb	PESQ-nb	STOI
UNet	Frequency	-26.99	18.11	-26.98	1.29	1.74	0.75
UNetDNP	Time	7.61	16.90	8.55	1.56	2.07	0.82
ConvTasNet	Time	10.69	22.19	11.30	1.83	2.38	0.85
TransUNet **	Frequency	-27.83	12.24	-27.81	1.42	1.90	0.75

Table 1. Objective evaluation results for all models.
Values represent the mean across all evaluation examples.

Conclusion: **Time domain** approaches give superior performance



Results: audio samples





Questions?

If you have questions, please send us an email

{hkt20, peggsr10, abbaslit10}@mails.tsinghua.edu.cn



Thank you
谢谢