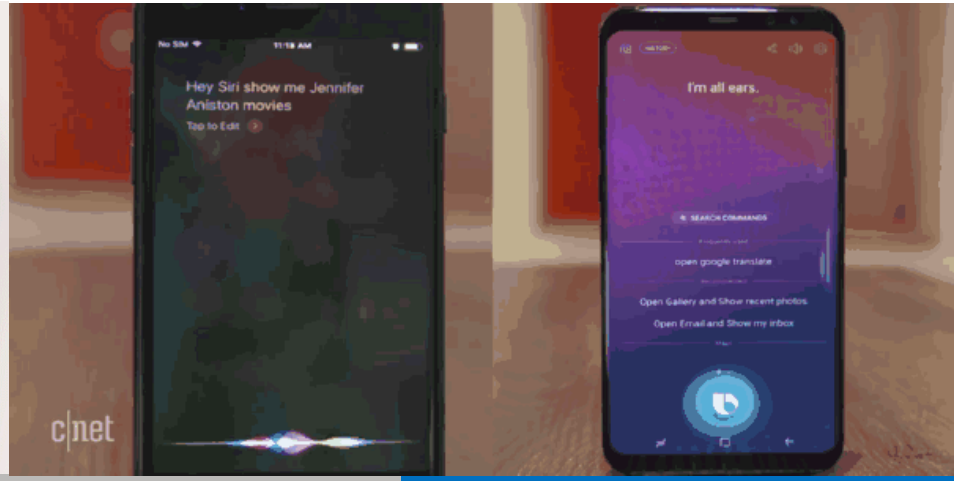
Research on Speech Separation Based on Deep Learning in Open Complex Environment

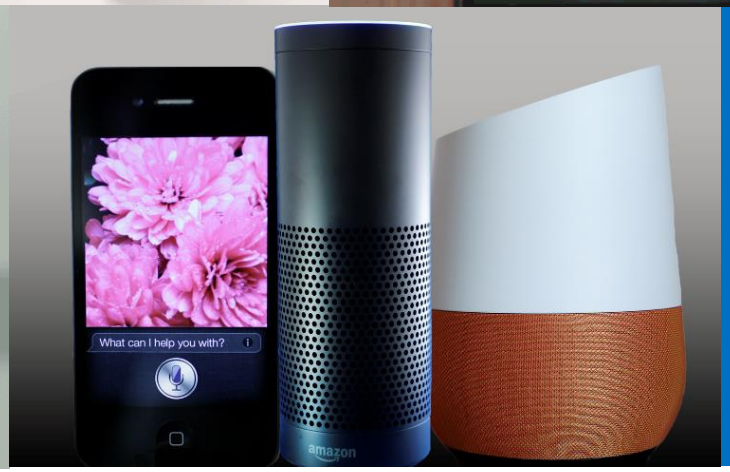
开放复杂环境下基于深度学习的语音分离方法研究

Reporter: Meng Ge

Supervisor: Longbiao Wang

Speech Applications

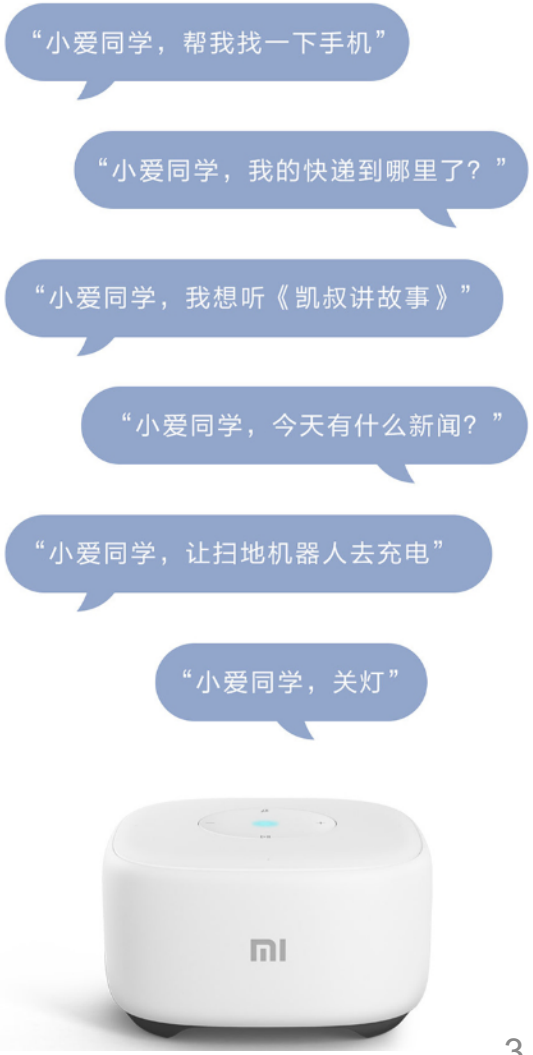




Applications

The watches, mobile phones, wristbands and voice boxes.

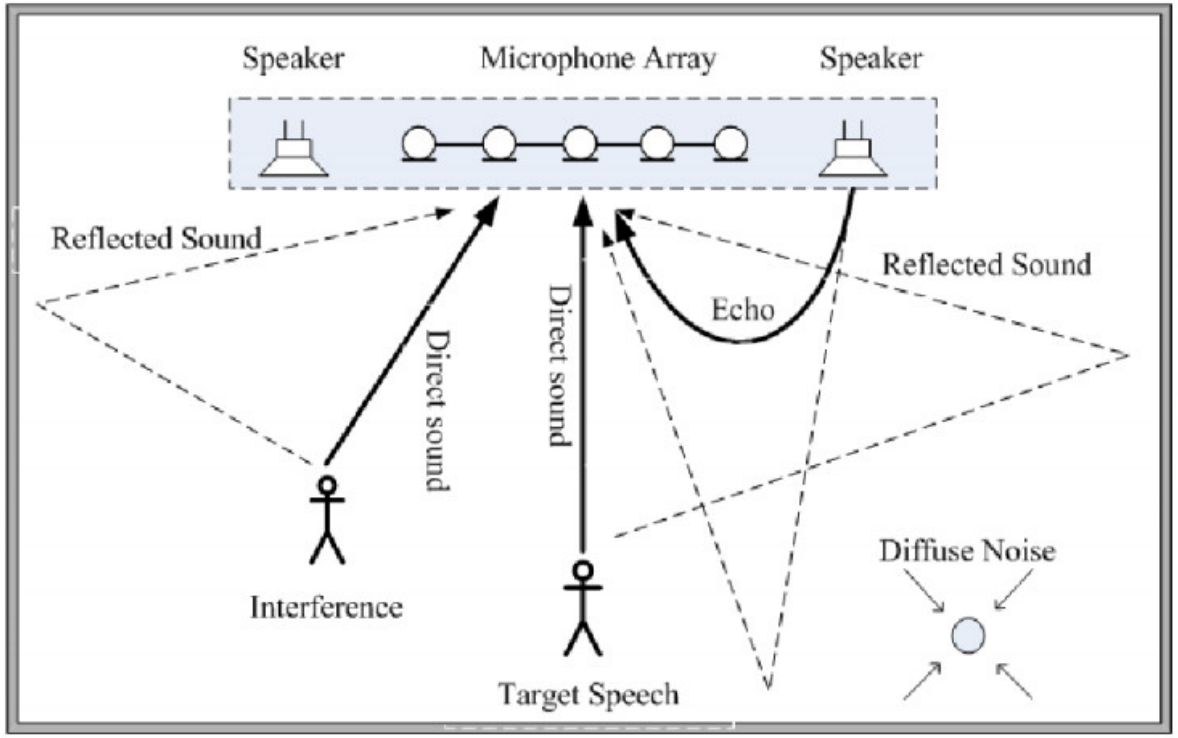
Speech Applications



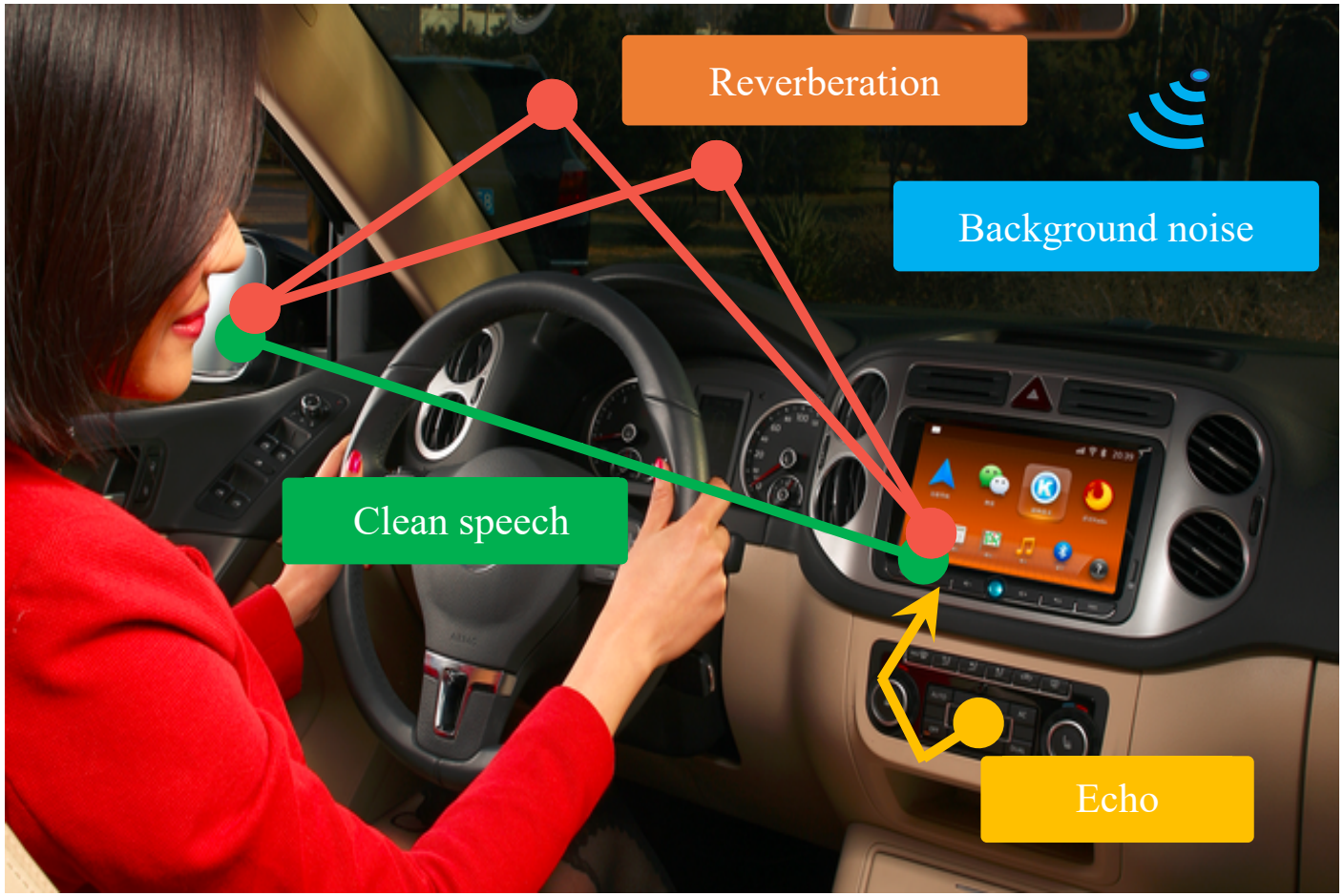
Problem: Cocktail Party Problem

- **Speech rarely occurs in isolation**
  - … but recognizing mixed speech is a problem

The cocktail party problem



Example: Car Environment



Clean Speech

It's the direct clean speech.

Reverberation

It's producted by reflection of the floor and windows.

Background noise

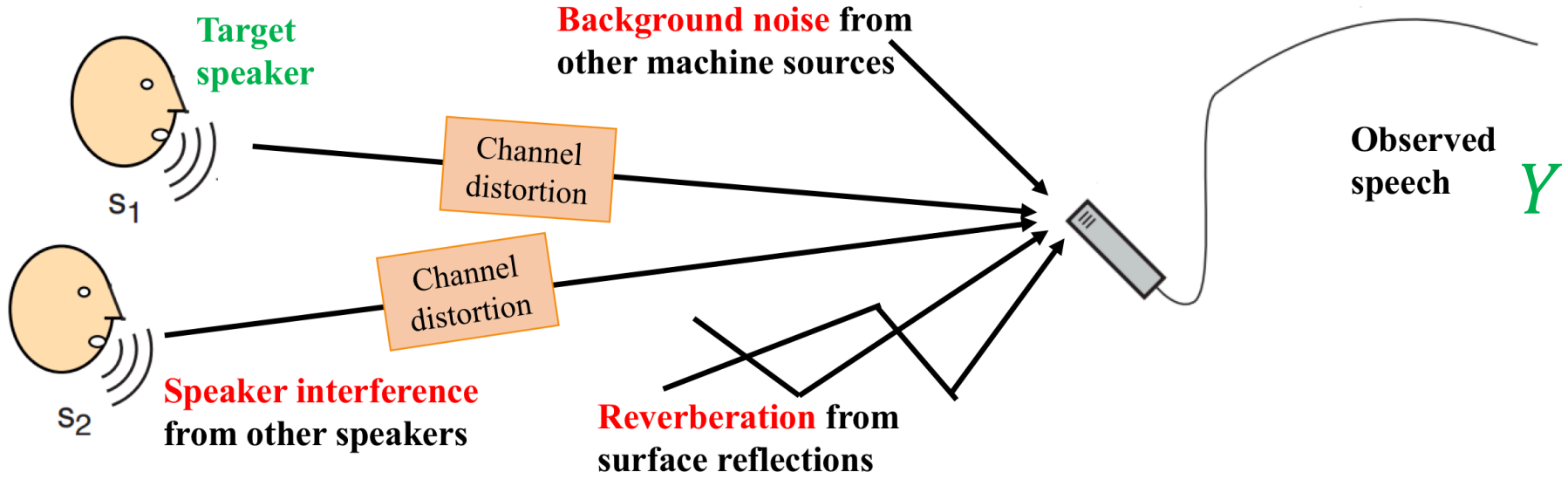
It contains some background speech, such as car noise.

Echo

It's a reflection of sound that arrives at the listener with a delay.

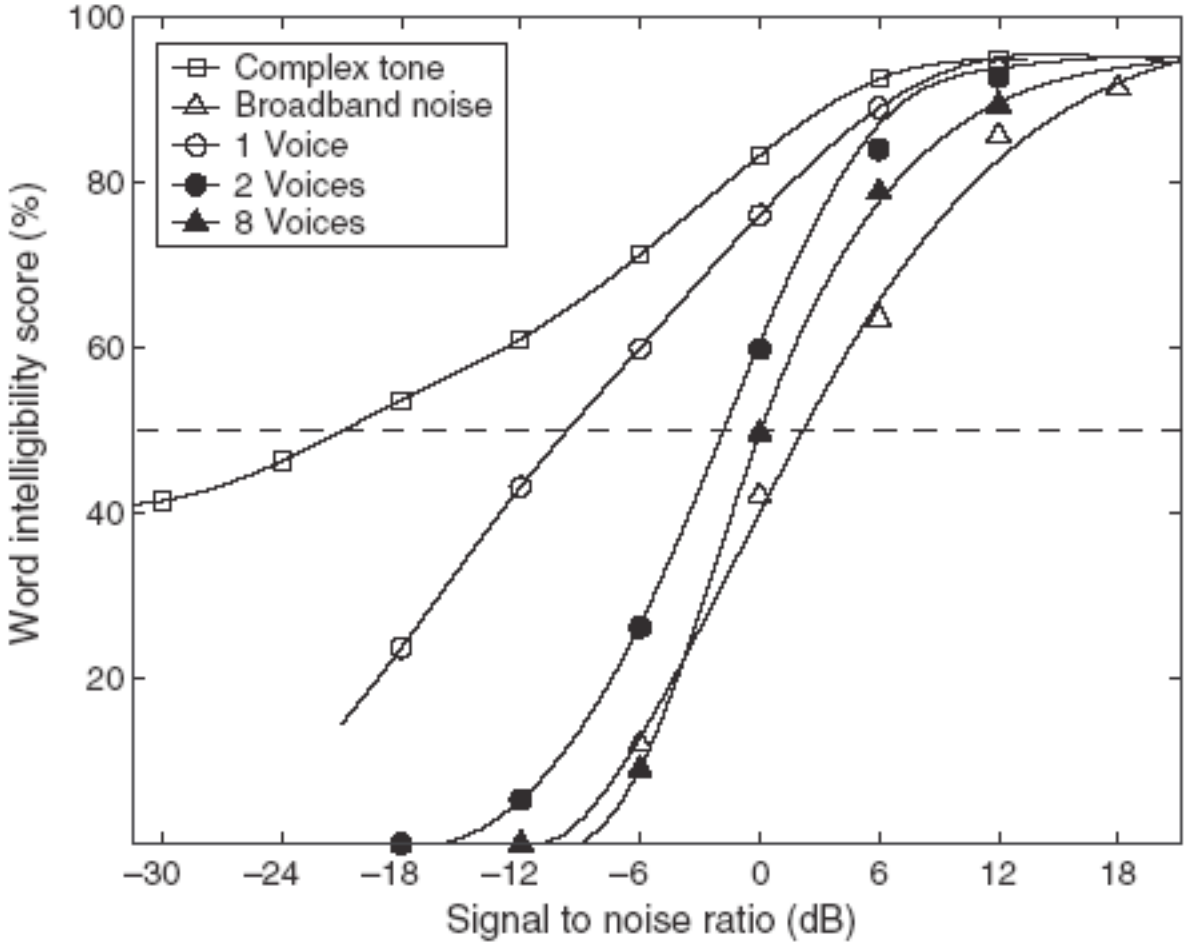
Definition: Cocktail Party Problem

- In open complex environment, speech rarely occurs in isolation. Target speech is always mixed with background interference, such as **speaker interference**, **reverberation** and **background noises**.



$$Y = \boldsymbol{s_1} * h + s_2 * h + N$$

Human Performance



SRT 23 dB difference in Speech Reception Threshold!

Source: Wang & Brown (2006)

Machine Solution: Speech Separation

- Speech separation is the task of separating target speech from background interference.
- According to the type of background interference, speech separation is usually divided into three categories:
  - **Speech enhancement (SE):** speech & non-speech separation
  - **Speech dereverberation (SD):** speech & reverberation separation
  - **Speaker separation (SS):** multi-talker separation

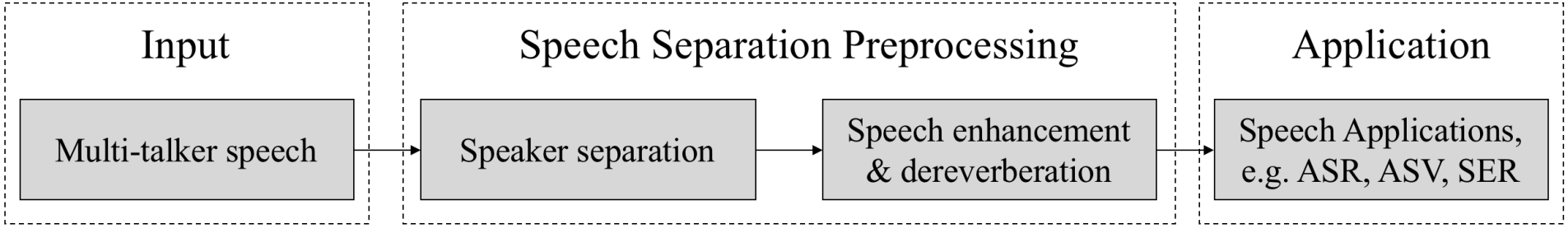
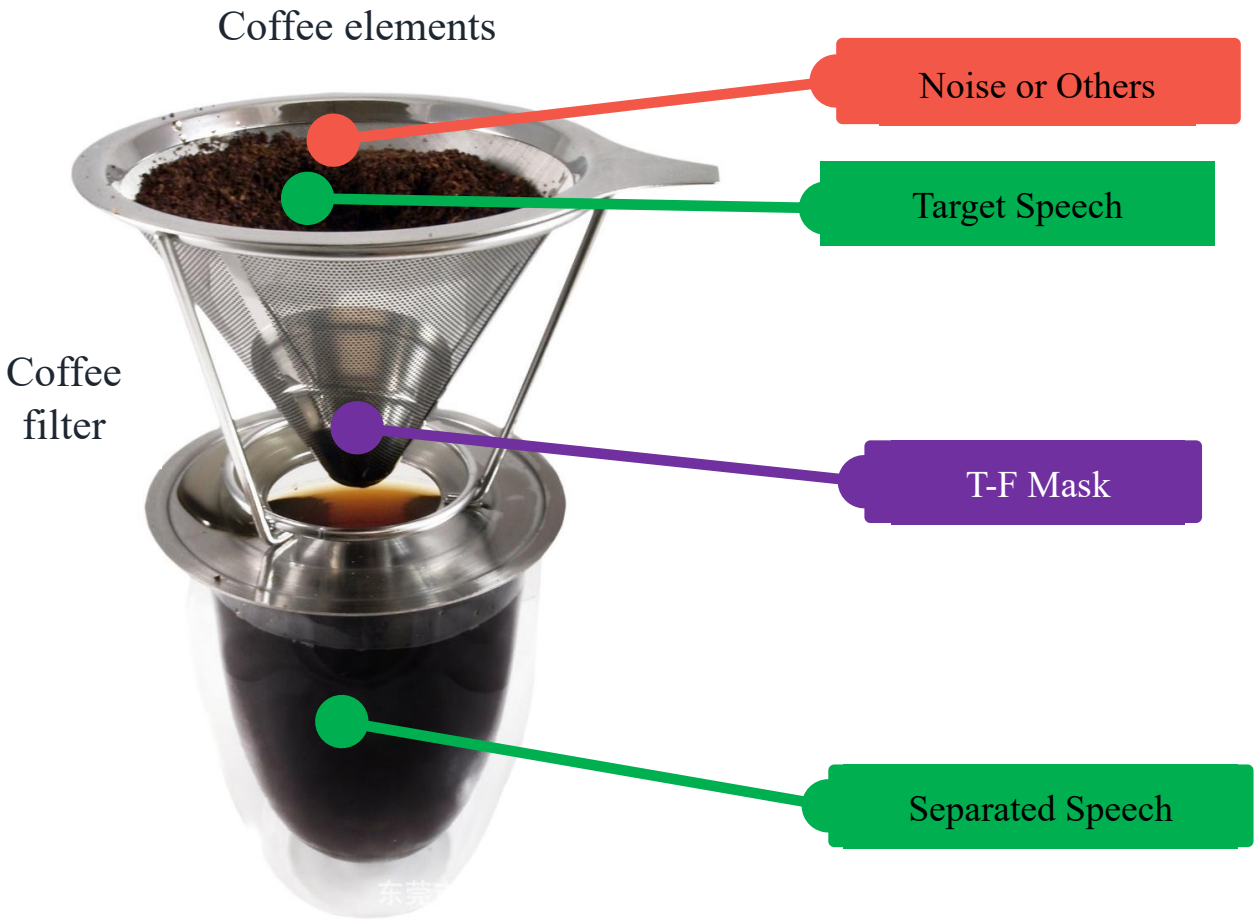


Figure 2. A flow chart of multi-talker speech separation for speech applications

Speaker Separation

- **Central idea: Which elements** and **How much** belong to the target source.



4	**5**	**3**
6	7	4
2	5	3

Mixture speech

0	**0.3**	**0.2**
0.2	0.8	0.6
0.4	0.2	1.0

Mask

0	**1.5**	**0.6**
1.2	5.6	2.4
0.8	1.0	3

Separated speech

General Framework: Single-channel

- General framework of single-channel speaker separation

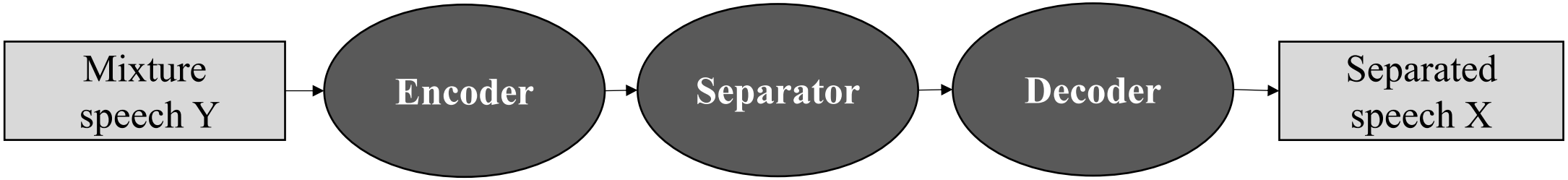


Figure 7. An illustration of speaker separation general framework

Model Type	Input Feature	Encoder/Decoder	Separation Network	Loss Function
Frequency-domain e.g. uPIT, TGT	Spectrogram	STFT/ISTFT	BLSTM	MSE
Time-domain e.g. TasNet, Wave-U-Net	Waveform	Conv-1D/ ConvTanspose-1D	CNN > BLSTM	Si-SNR or MAE

Table 1. The details of two speaker separation framework: frequency-domain & time-domain

General Framework: Single-channel

- The mainstream methods are divided into 4 categories:
  1) **Deep Clustering (DPCL)**; 2) **PIT**; 3) **CASA**; 4) **Speaker Extraction**

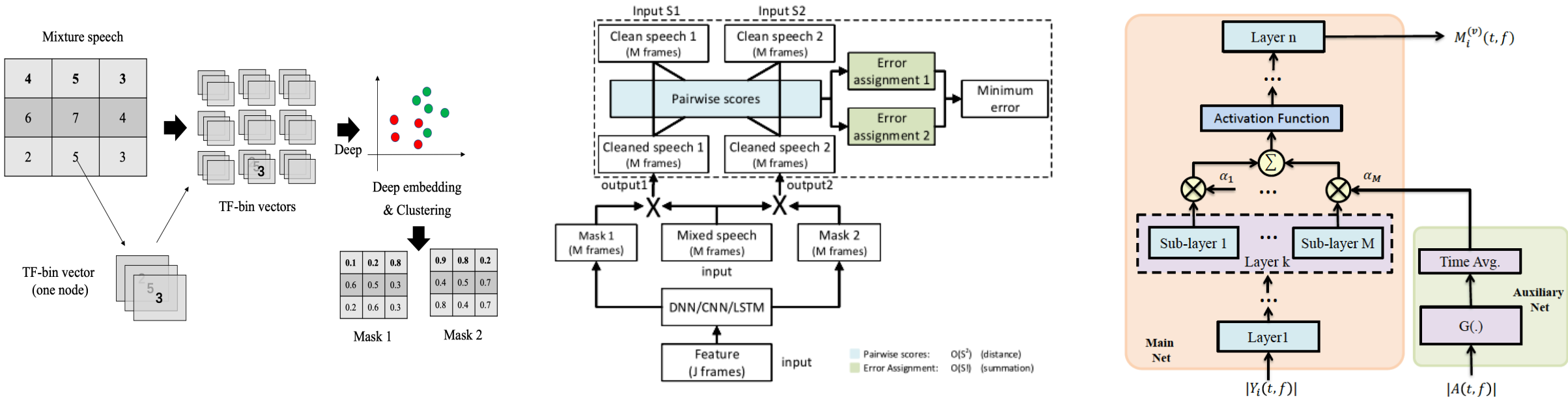


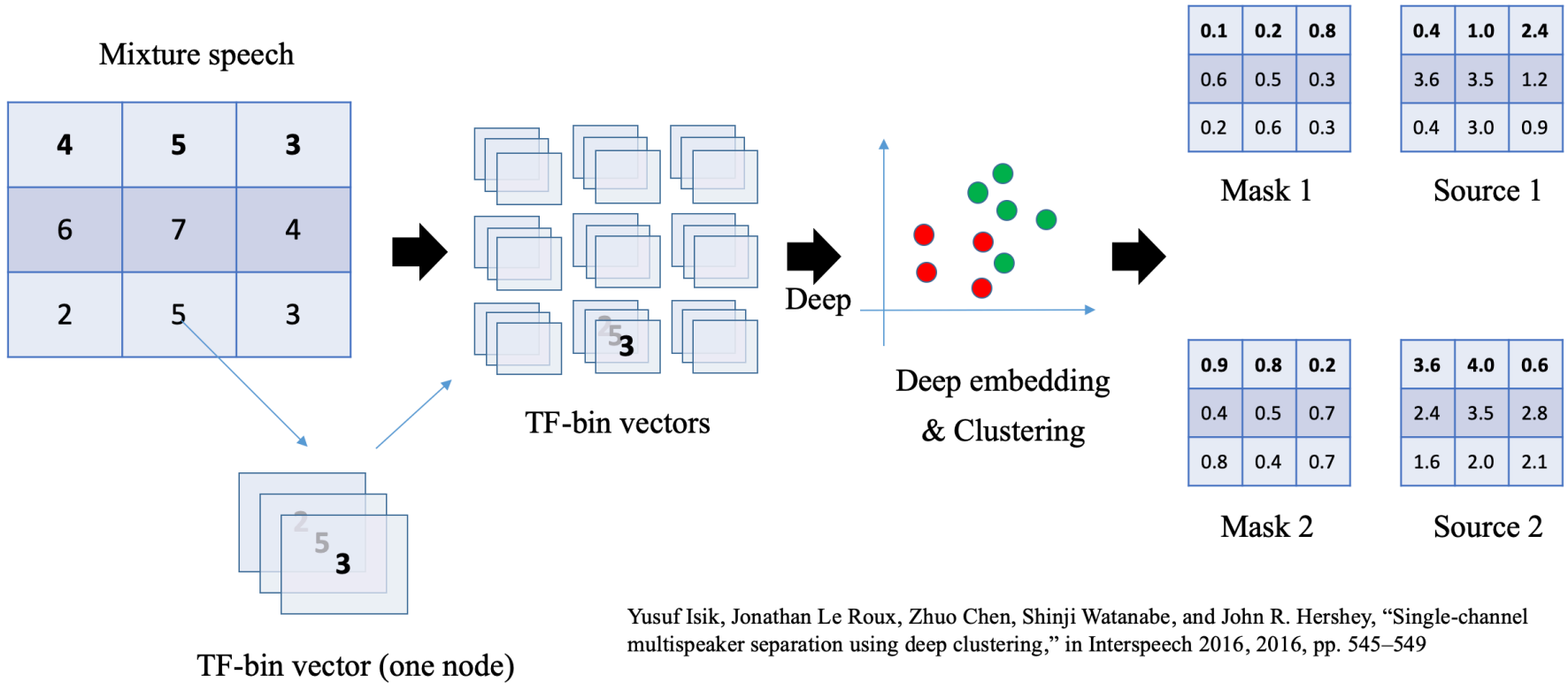
Figure 8. An illustration of classical DPCL (Left), PIT model (Middle) and Target speaker extraction frameworks (Right)

Problems in speaker separation

- Permutation Problem
- Speaker Number Mismatch

Deep Clustering (DPCL)

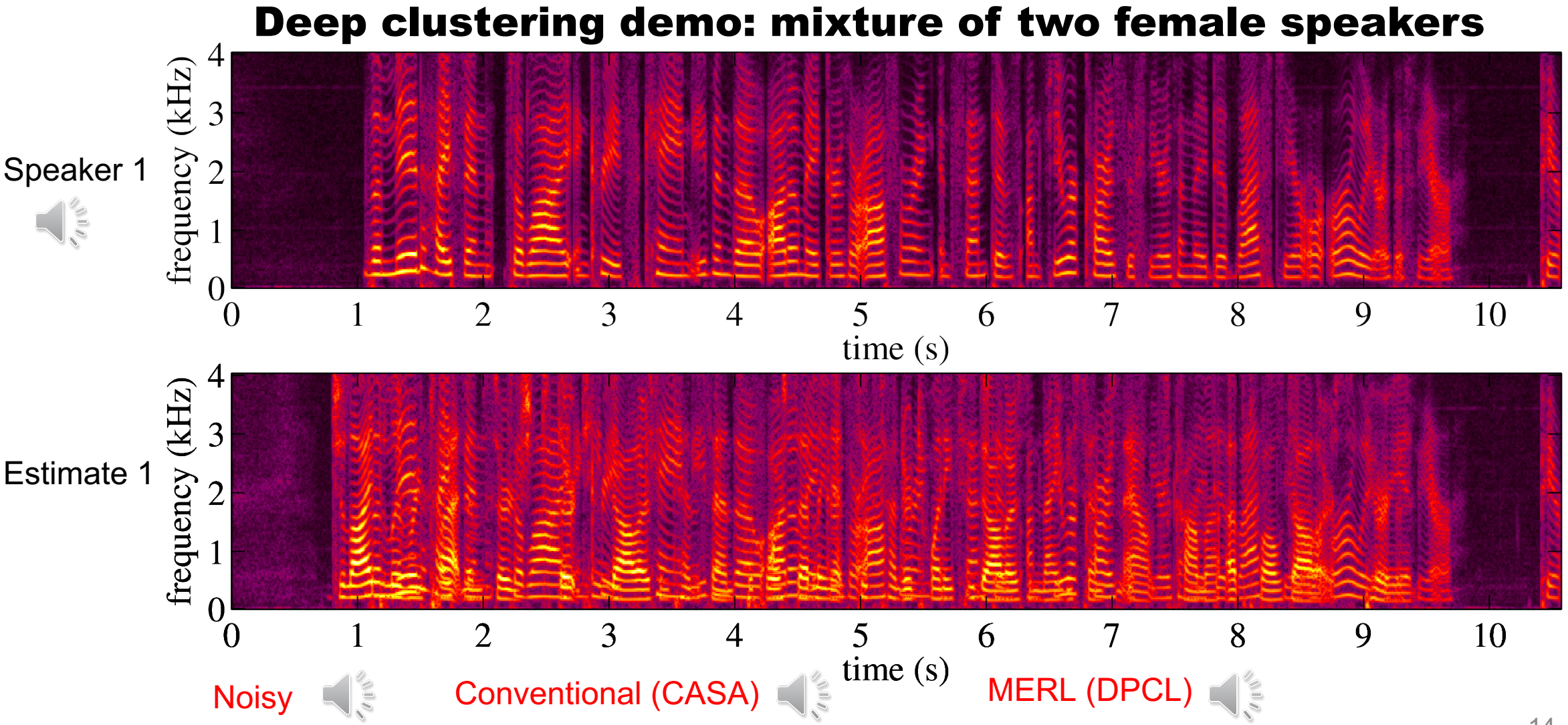
- Deep clustering method considers speaker separation problem as a clustering problem, which estimates mask using clustering strategy.



Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R. Hershey, "Single-channel multispeaker separation using deep clustering," in Interspeech 2016, 2016, pp. 545–549

Website: https://scholar.google.com/citations?hl=zh-CN&user=ACcAdXgAAAAJ&view_op=list_works

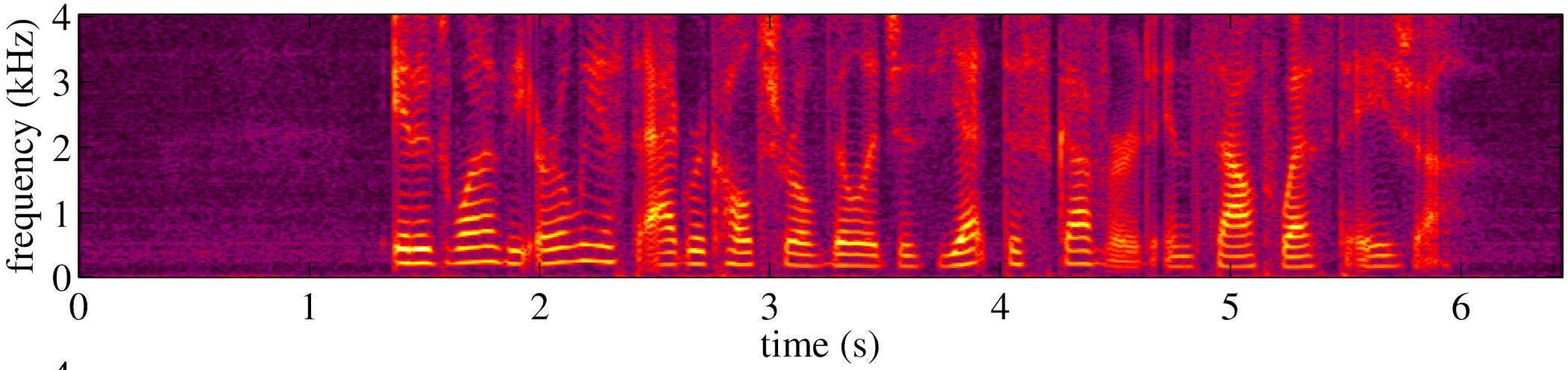
Deep Clustering (DPCL) Demo



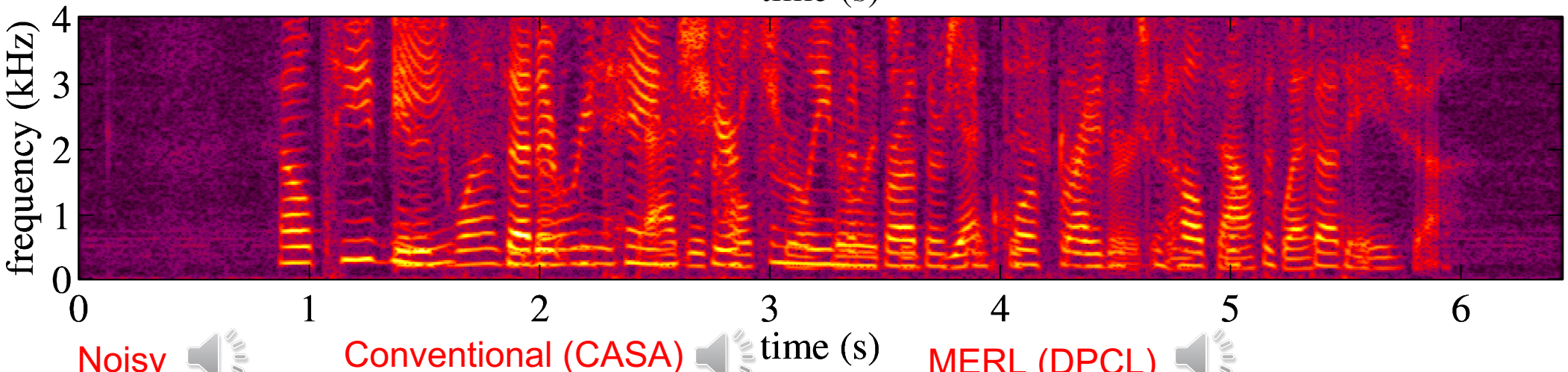
Deep Clustering (DPCL) Demo

Deep clustering demo: mixture of two female speakers, “failing”

Speaker 1



Estimate 1



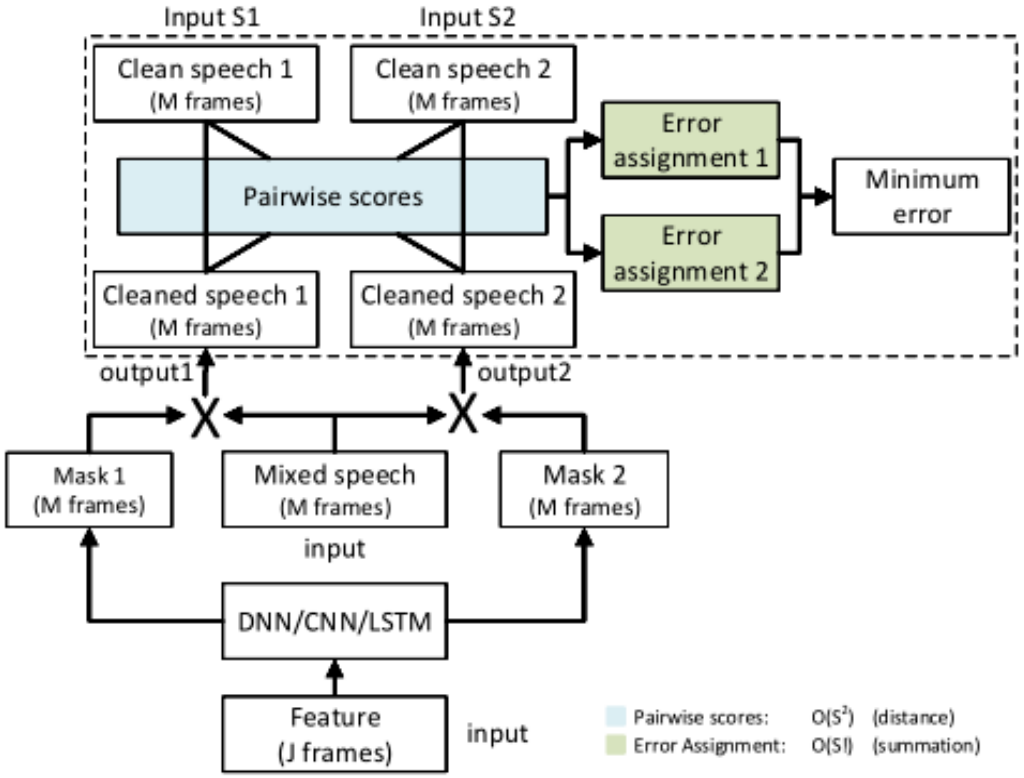
Noisy

Conventional (CASA)

MERL (DPCL)

PIT

- The strategy of using mathematical permutation and combination is utilized to solve the permutation/label problem.



$$L_1 = \|\hat{s}_1 - s_1\|^2 \|\hat{s}_2 - s_2\|^2$$

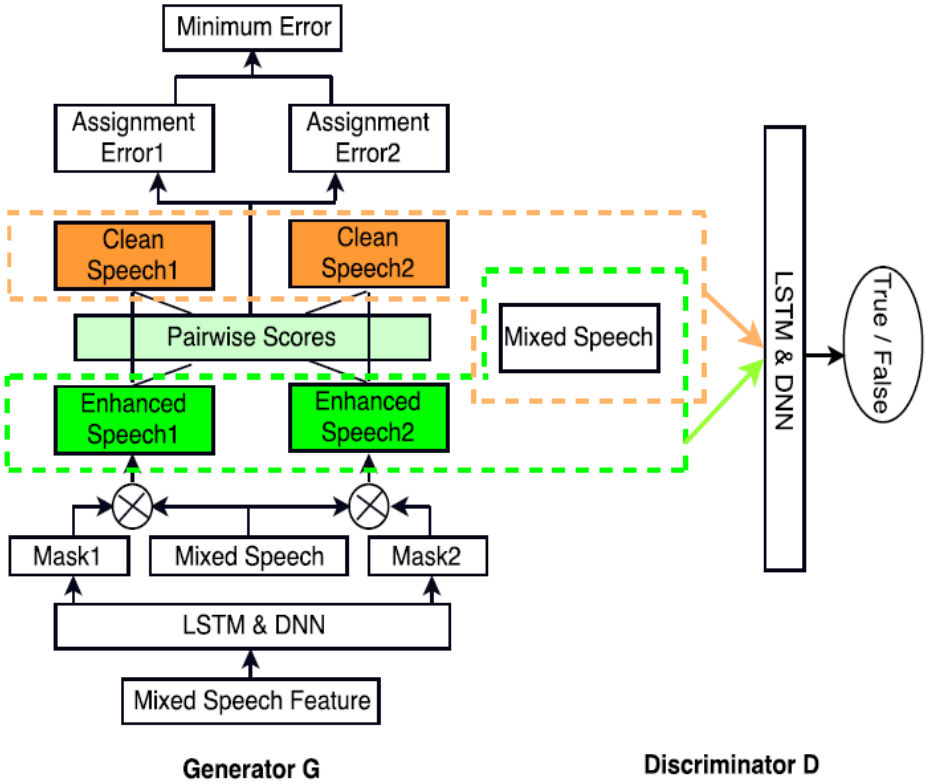
$$L_2 = \|\hat{s}_1 - s_2\|^2 \|\hat{s}_2 - s_1\|^2$$

$$L = \min(L_1, L_2)$$

Website: https://scholar.google.com/citations?user=tMY31_gAAAAJ&hl=zh-CN&oi=sra

SSGAN-PIT

- Problem: The drawbacks of **MSE loss.** (Perceptual discrepancies)
- Solution: Use the discriminator of GAN instead of MSE



$$\min_D V_{LSGAN}(D) = \frac{1}{2}\mathbb{E}_{Y,X_s\sim p_{data}}[(D(|Y|,|X_{s=1:S}|)-1)^2] + \frac{1}{2}\mathbb{E}_{Y\sim p_{data}}[D(|Y|,G(log|Y|^2))^2] \quad (4)$$

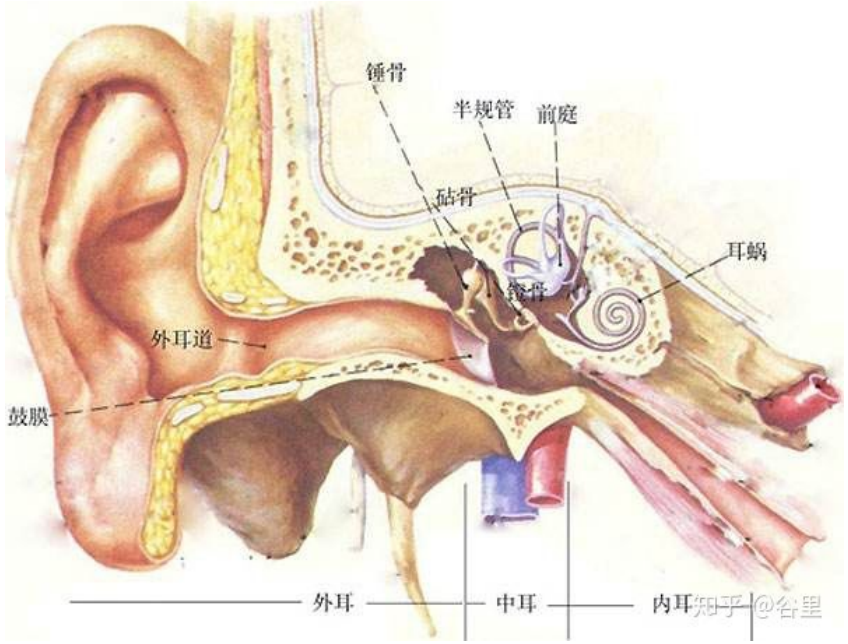
$$\min_G V_{LSGAN}(G) = \frac{\lambda}{2}\mathbb{E}_{Y\sim p_{data}}[(D(|Y|,G(log|Y|^2))-1)^2] \quad (5)$$

Core solution $+\mathcal{J}_{SS}$

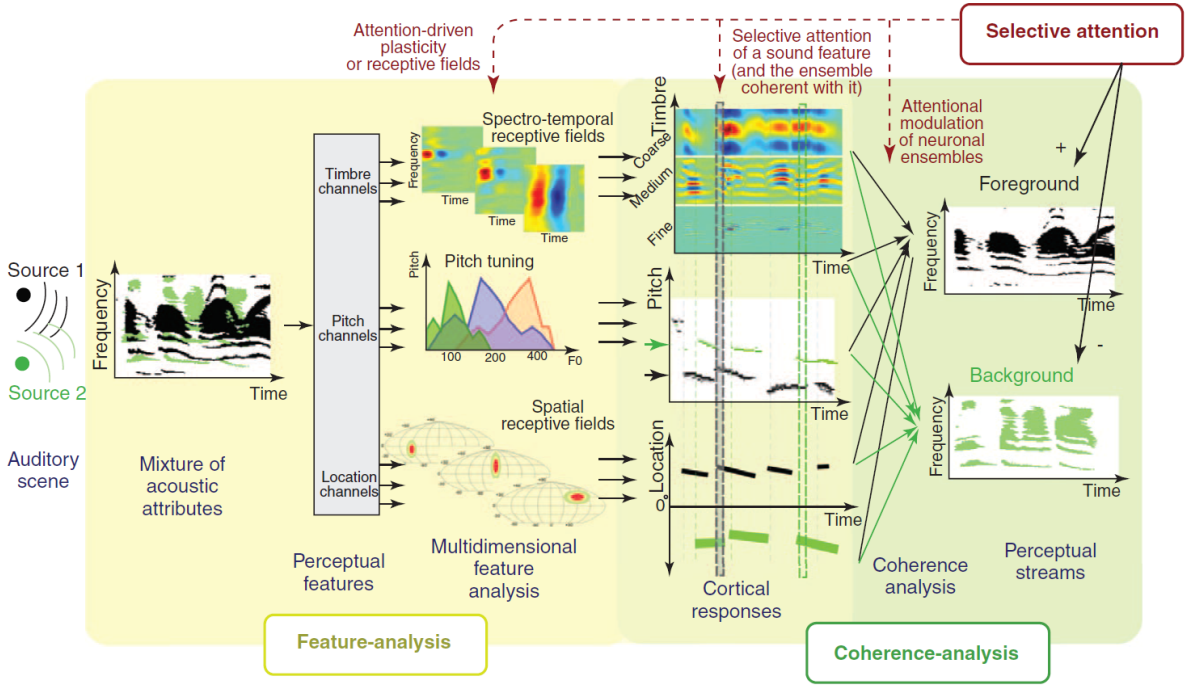
$$\mathcal{J}_{SS} = \frac{1}{T\times F\times S}\sum_{s=1}^{S} \| \hat{M}_s \otimes |Y| - |X_s| \|_F^2,$$

CASA

- **Auditory Scene Analysis** (**ASA**) is a proposed model for the basis of auditory perception. This is understood as the process by which the human auditory system organizes sound into perceptually meaningful elements.



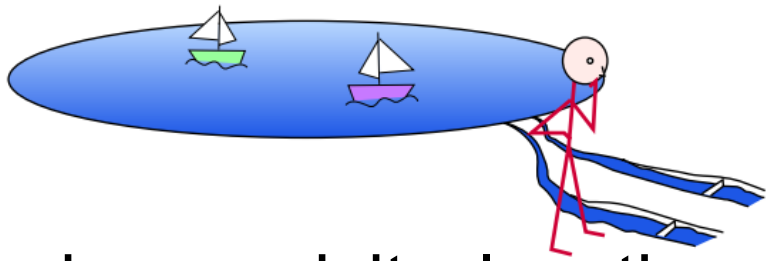
Auditory Pathway



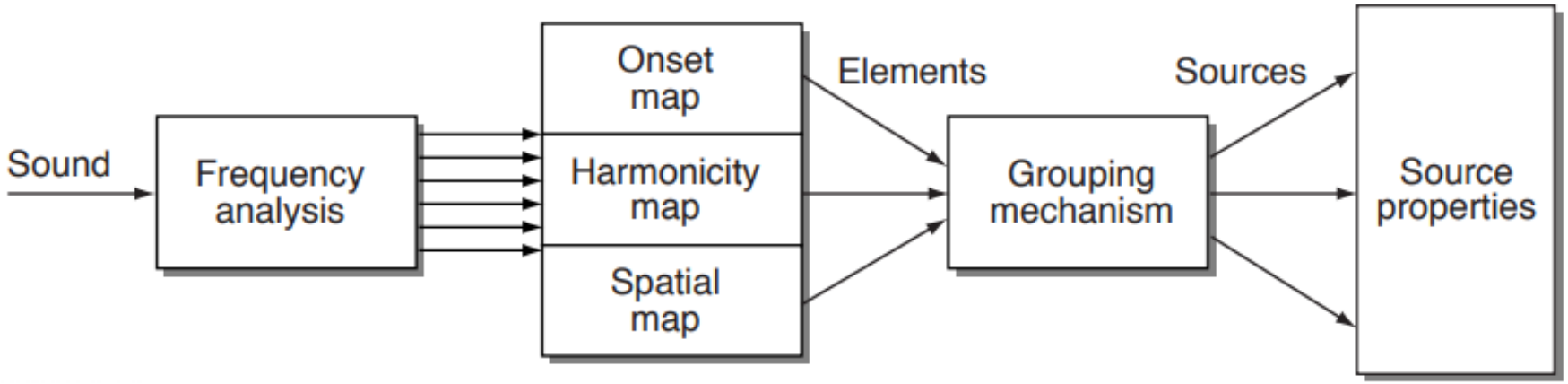
ASA (Author:Psychologist Albert Bregman)
(初级分析+图式加工)

CASA

- **How do people analyze sound mixtures?**
  - break mixture into small elements (in time-frequency)
  - elements are grouped into sources using cues
  - sources have aggregate attributes

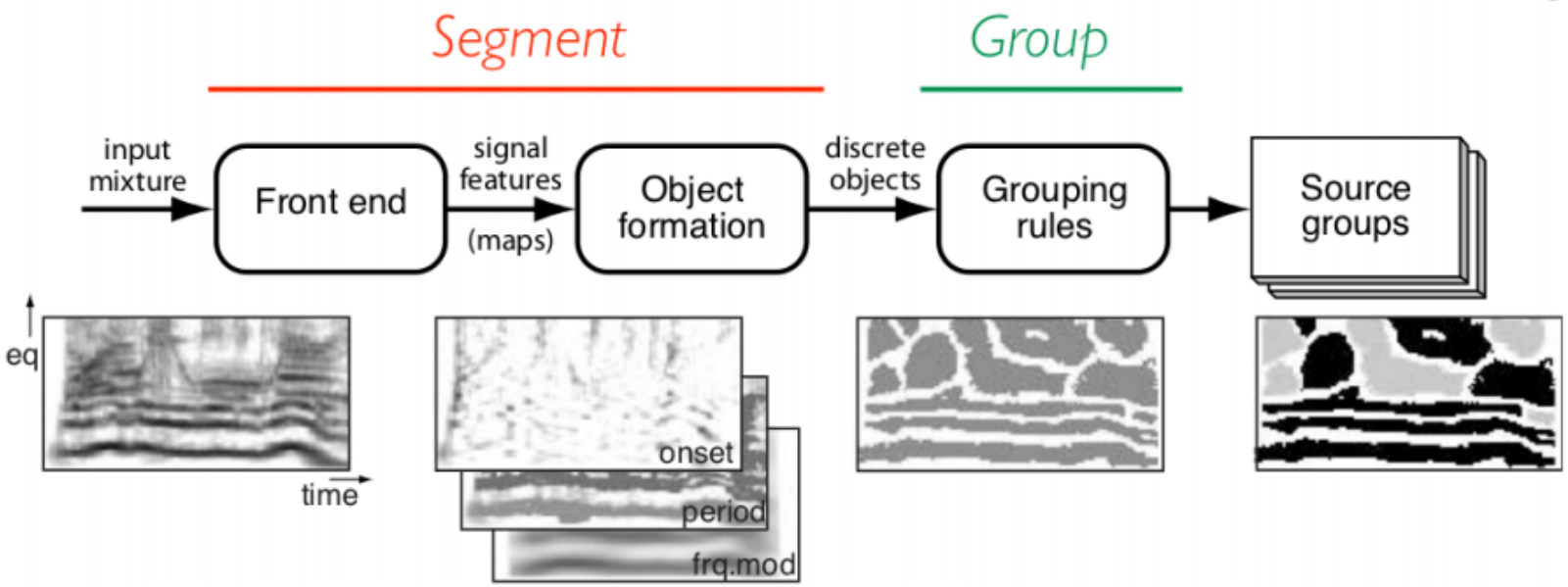


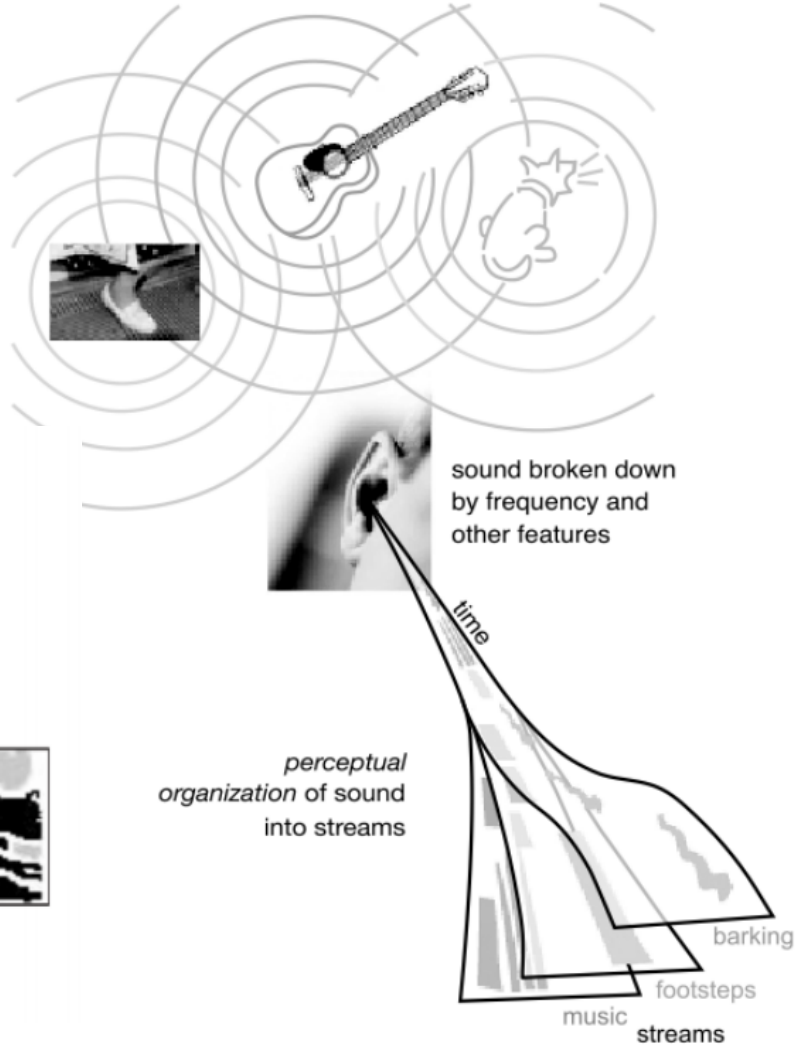
- **Grouping rules:**
  - cues: common onset/offset/modulation, harmonicity, location,…



CASA

- **Computer system for separating sounds**
  - based on biological "inspiration" (ASA)
  - Segment time-frequency into sources based on perceptual grouping cues



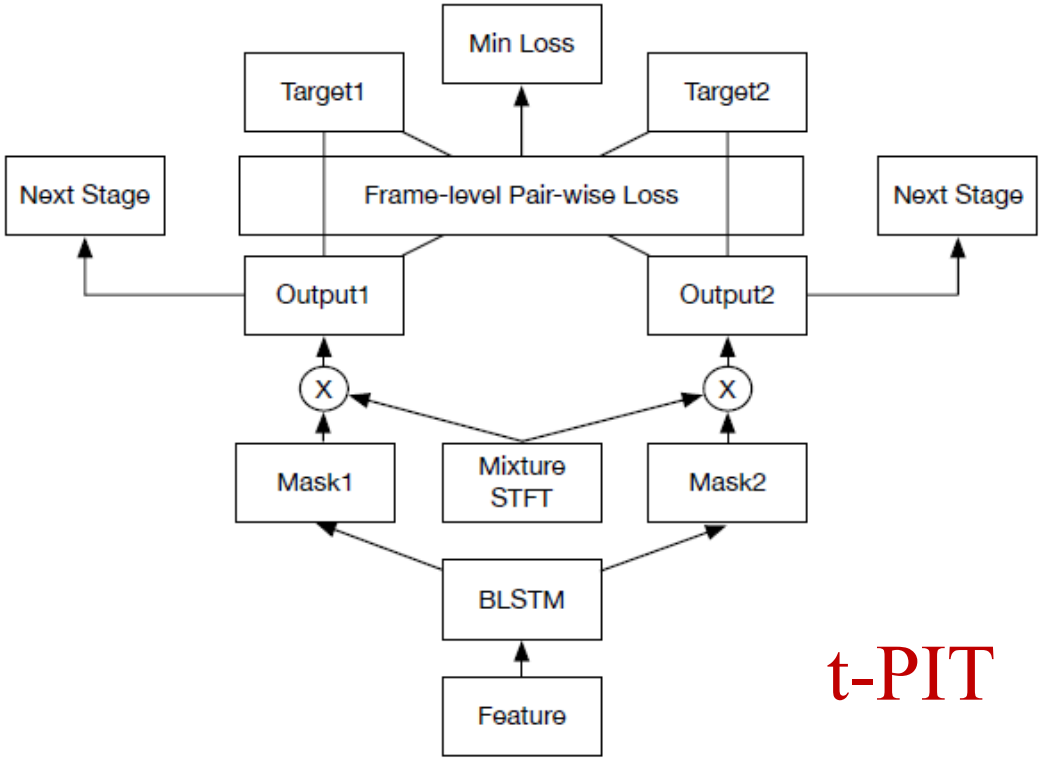


CASA & PIT

- CASA is similar to PIT framework (Mask-inference)
- Now: CASA idea + PIT idea
- Difference: 1. features; 2. mechanism (freq + Time)

CASA

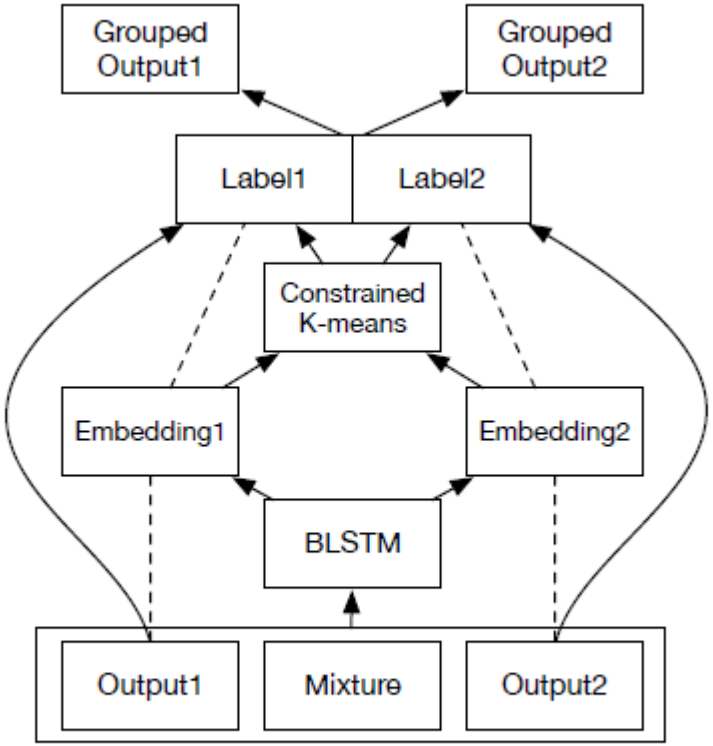
Frequency-domain



t-PIT

Fig. 1: Diagram of the simultaneous grouping stage.

Time-domain



DPCL

Fig. 2: Diagram of the sequential grouping stage.

Yuzhou Liu and DeLiangWang, "A CASA Approach to Deep Learning Based Speaker-Independent Co-Channel Speech Separation," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5399–5403.

Chimera Network

DPCL Mask-inference

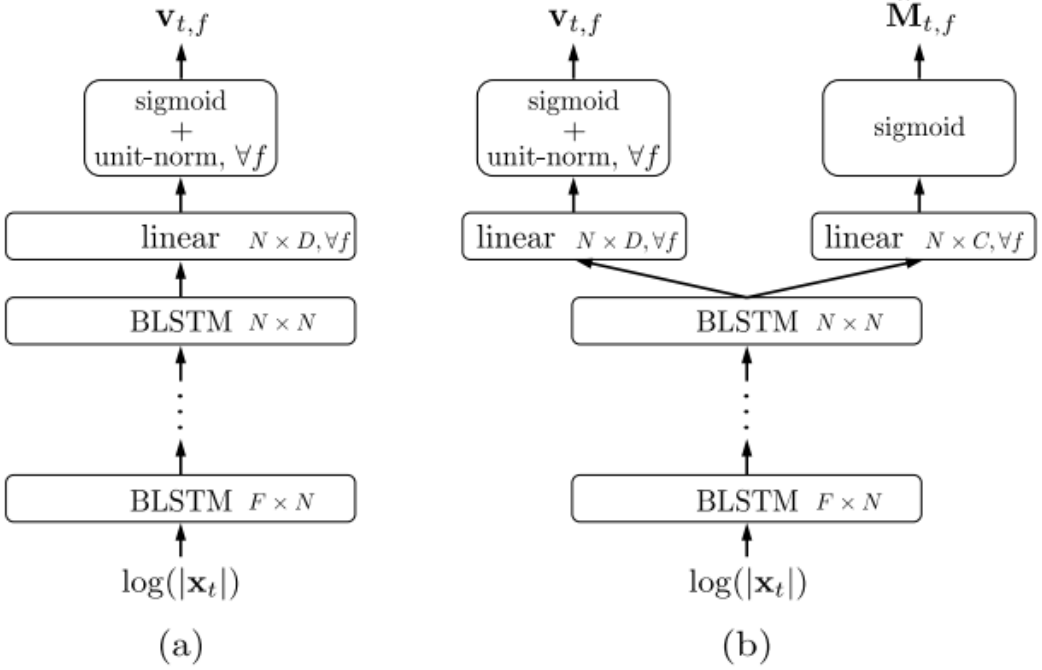


Fig. 1. (a) Deep clustering network, (b) Chimera++ network

- Q: Why the multi-task structure is better? ([1] Sec. 2.2)

  A: Whereas conventional mask-inference approaches only focus on increasing the separation between sources, the deep clustering objective also reduces within-source variance in the internal representation, which could be beneficial for generalization.

- Objective function

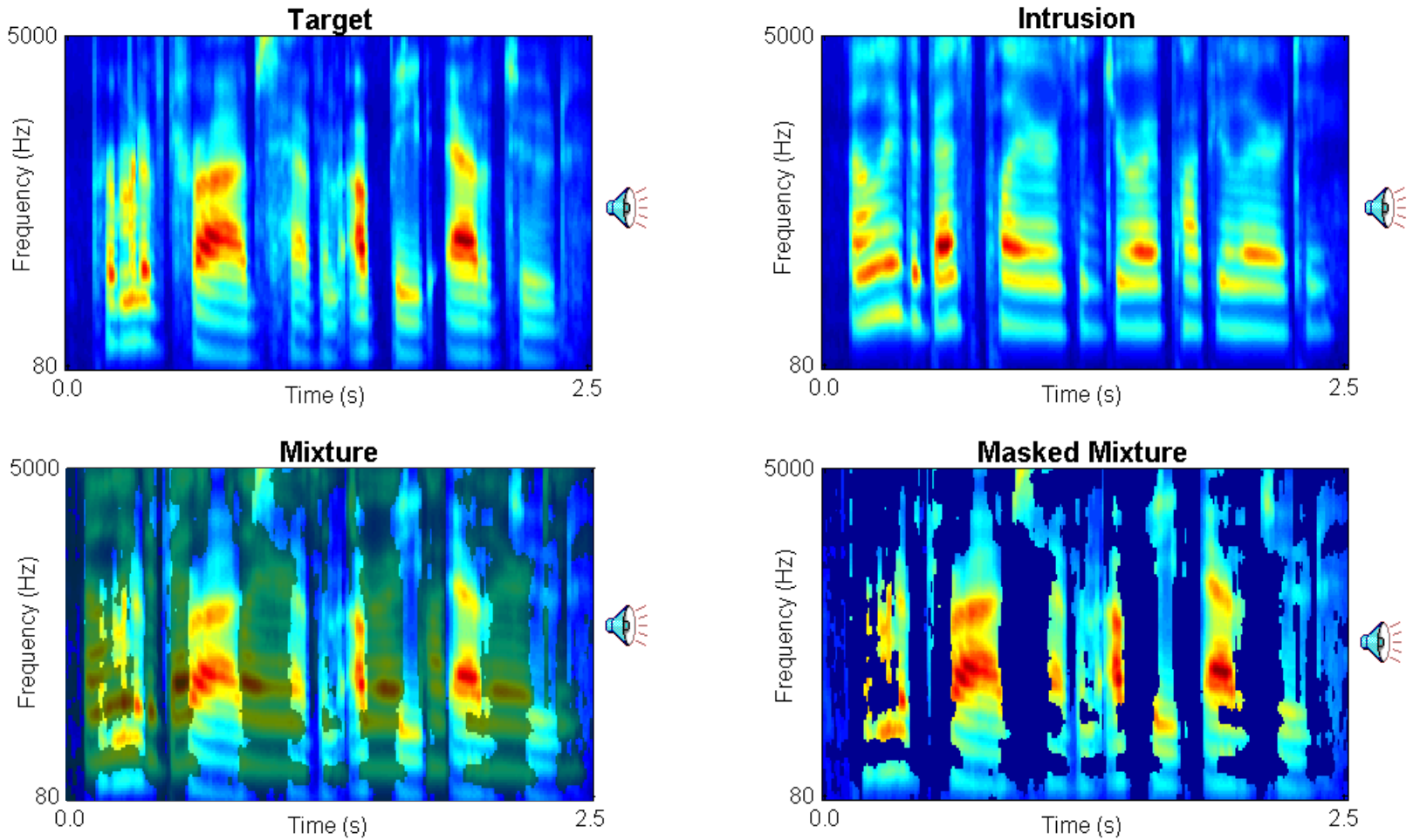
$$\mathcal{L}_{\text{chi}^{++}_{\alpha}} = \alpha \mathcal{L}_{\text{DC}}(V, Y) + (1 - \alpha)\mathcal{L}_{\text{MI}}.$$

[1] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep Clustering and Conventional Networks for Music Separation: Stronger Together," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Mar. 2017.

[2] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey, "Alternative Objective Functions for Deep Clustering," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 686–690.

[3] Zhong-Qiu Wang, Jonathan Le Roux, DeLiang Wang, and John R. Hershey, "End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction," in Interspeech 2018, 2018, pp. 2708–2712.

Highlight : Mask



Highlight : Mask

- **TBM (Kjems et al.'09; Gonzalez & Brookes'14) is similar to the IBM except that interference is fixed to speech-shaped noise (SSN)**
- **IRM (Srinivasan et al.'06; Narayanan & Wang'13; Wang et al.'14; Hummersone et al.'14)**

$$IRM(t,f) = \left(\frac{S^2(t,f)}{S^2(t,f) + N^2(t,f)} \right)^{\beta} = \left(\frac{SNR(t,f)}{SNR(t,f)+1} \right)^{\beta}$$

- S and N denote speech and noise
- β is a tunable parameter, and a good choice is 0.5
- With $\beta = 0.5$, the IRM becomes a square root Wiener filter, which is the optimal estimator of the power spectrum

Highlight : Mask

- **Spectral magnitude mask (Wang et al.'14)**

$$SMM(t,f) = \frac{|S(t,f)|}{|Y(t,f)|}$$

  - Y denotes noisy signal

- **Phase-sensitive mask (Erdogan et al.'15)**

$$PSM(t,f) = \frac{|S(t,f)|}{|Y(t,f)|}\cos\theta$$

  - θ denotes the difference of the clean speech phase and noisy speech phase within the T-F unit
  - Because of phase sensitivity, this target usually leads to a better estimate of clean speech than the SMM mask

Highlight : Mask

- **This mask is defined so that, when applied, it results in clean speech (Williamson et al.'16)**

$$S(t,f) = cIRM * Y(t,f)$$

- **With complex numbers, solve for mask components**

$$cIRM(t,f) = \frac{Y_r S_r + Y_i S_i}{{Y_r}^2 + {Y_i}^2} + i\frac{Y_r S_i - Y_i S_r}{{Y_r}^2 + {Y_i}^2}$$

  - Subscripts *r* and *i* denote real and imaginary components
  - Some form of compression (e.g. tangent hyperbolic function) should be used to bound mask values

Highlight : Mask

- **Target magnitude spectrum (TMS) (Lu et al.'13; Xu et al.'14; Han et al.'14)**

$$|S(t,f)|$$

  - A common form of the TMS is the log-power spectrum of clean speech

- **Gammatone frequency target power spectrum (GF-TPS) (Wang et al.'14)**

$$S_{GF}^2(t,f)$$

- **The estimation of these two targets corresponds to spectral mapping, as opposed to T-F masking for earlier targets**

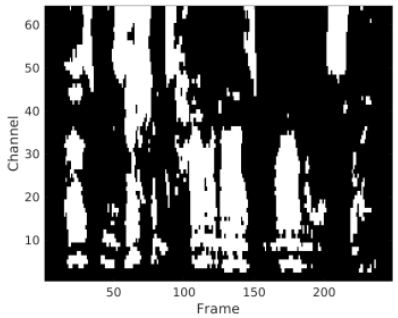
Highlight : Mask

- **In signal approximation (SA), training aims to estimate the IRM but the error is measured against the spectral magnitude of clean speech (Weninger et al.'14)**

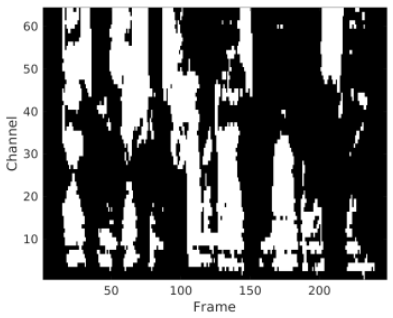
$$SA(t,f) = (RM(t,f)|Y(t,f)| - |S(t,f)|)^2$$

  - $RM(t,f)$ denotes an estimated IRM
  - This objective function maximizes SNR

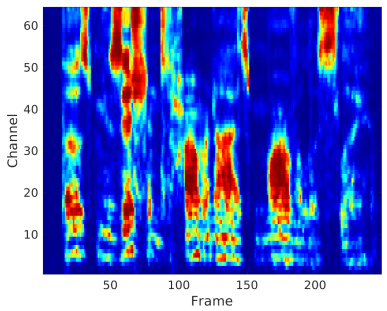
Highlight : Mask



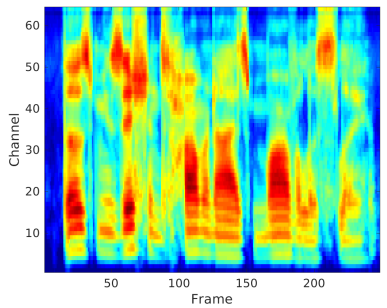
(a) IBM



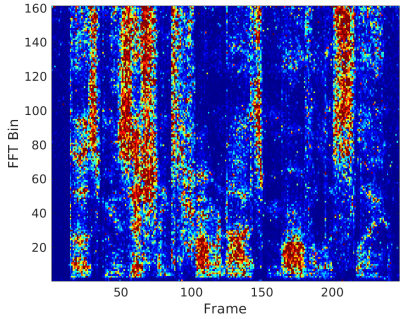
(b) TBM



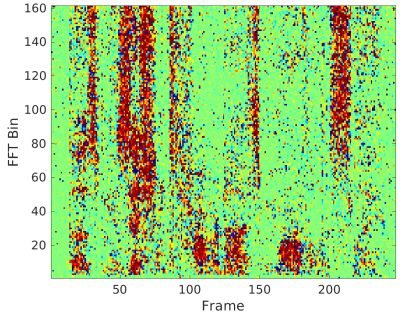
(c) IRM



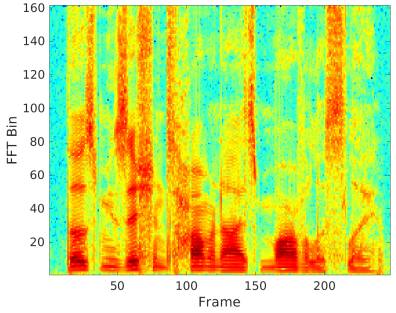
(d) GF-TPS



(e) SMM



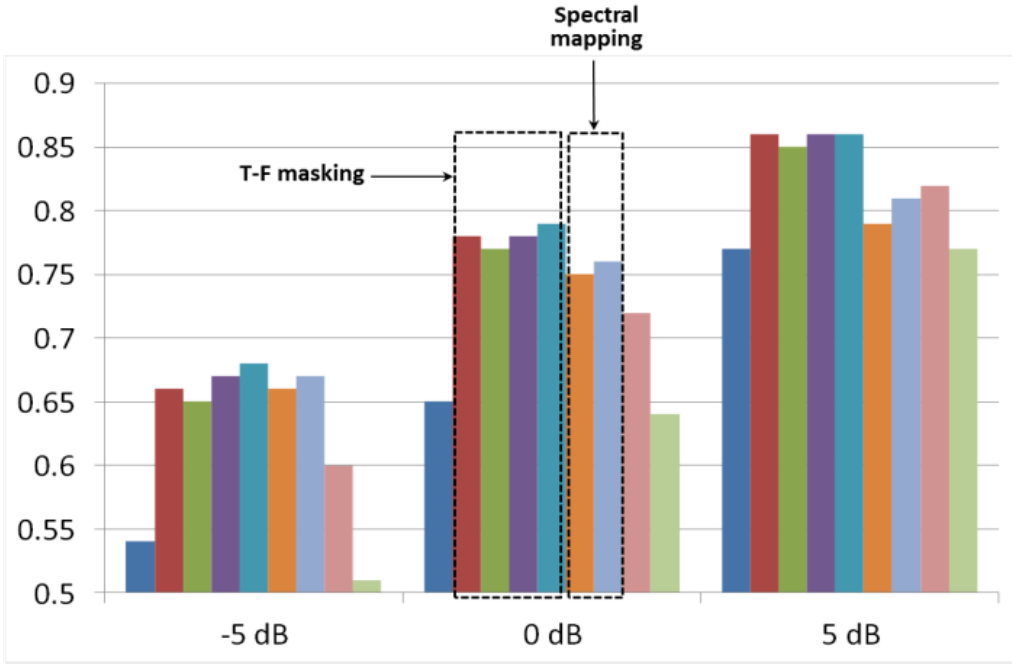
(f) PSM



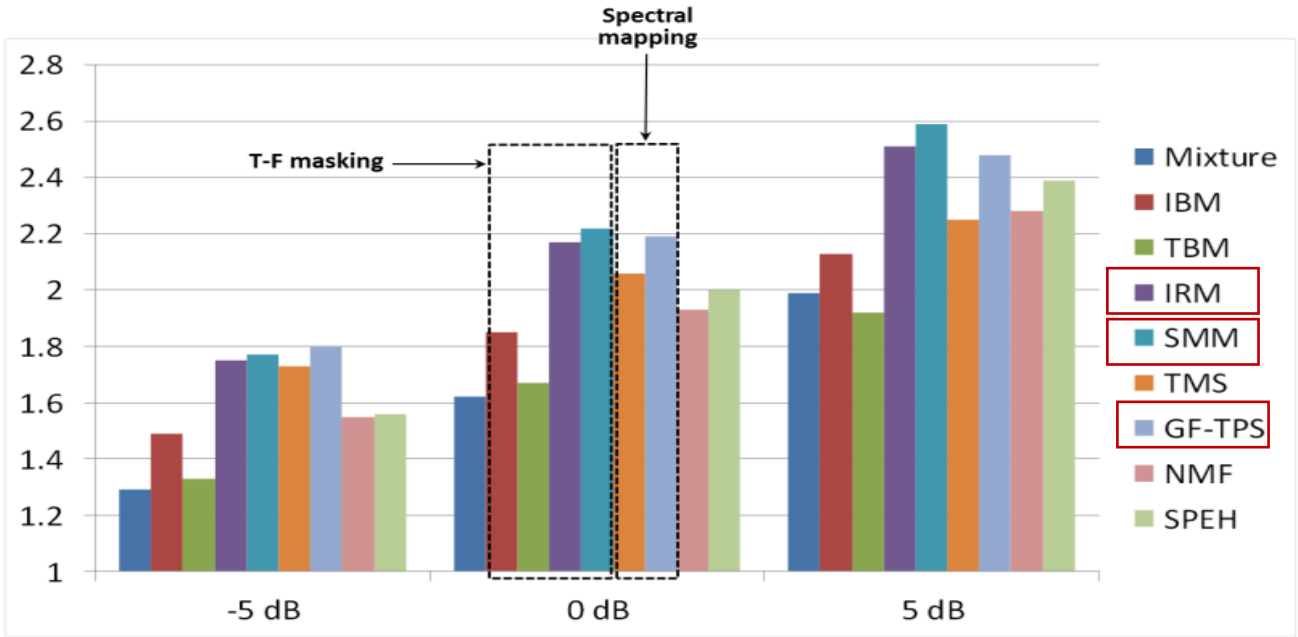
(g) TMS

Factory noise at -5 dB

Highlight : Mask



STOI comparison for factory noise



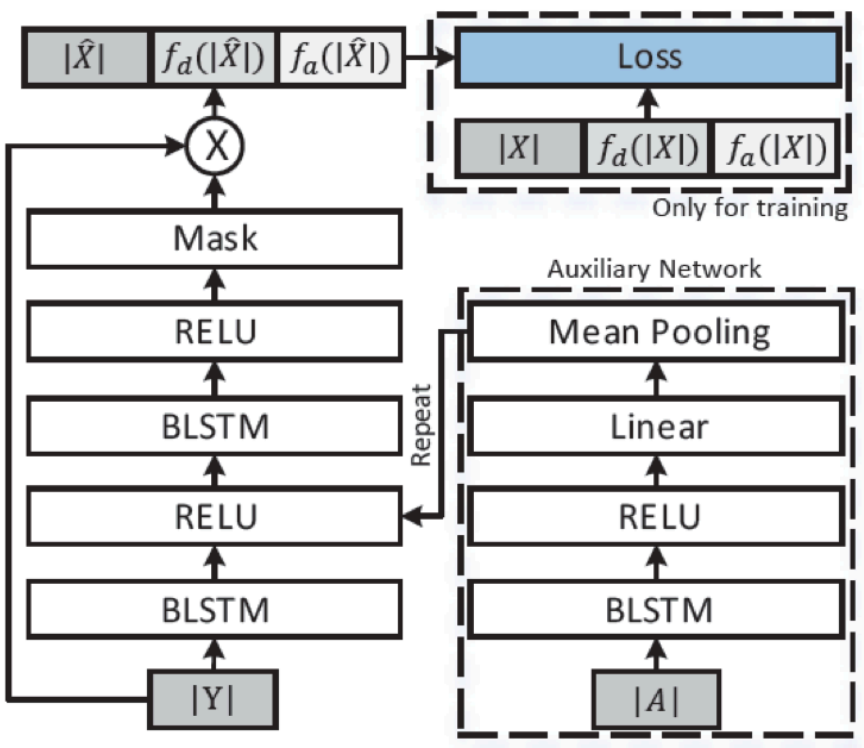
PESQ comparison for factory noise

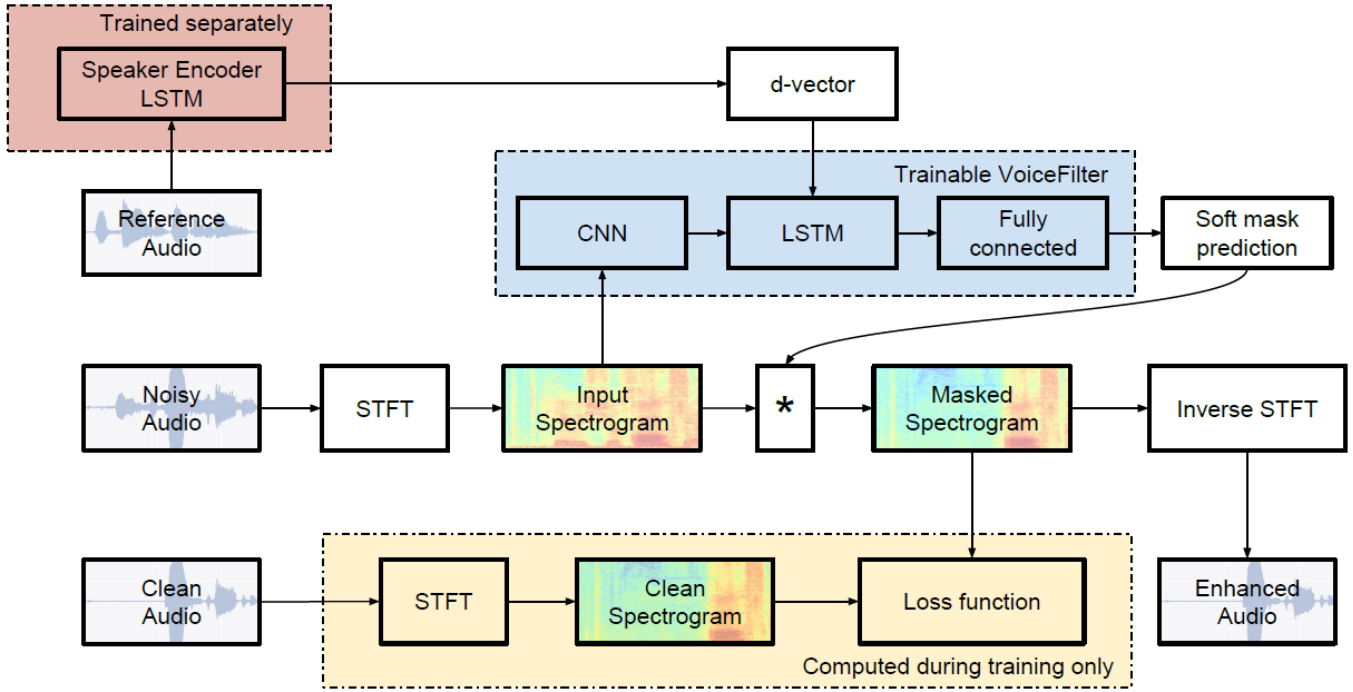
Highlight : Mask

- **Among the two binary masks, IBM estimation performs better in PESQ than TBM estimation**
- **Ratio masking performs better than binary masking for speech quality**
  - IRM, SMM, and GF-TPS produce comparable PESQ results
- **SMM is better than TMS for estimation**
  - Many-to-one mapping in TMS vs. one-to-one mapping in SMM, and the latter should be easier to learn
  - Estimation of spectral magnitudes or their compressed version tends to magnify estimation errors

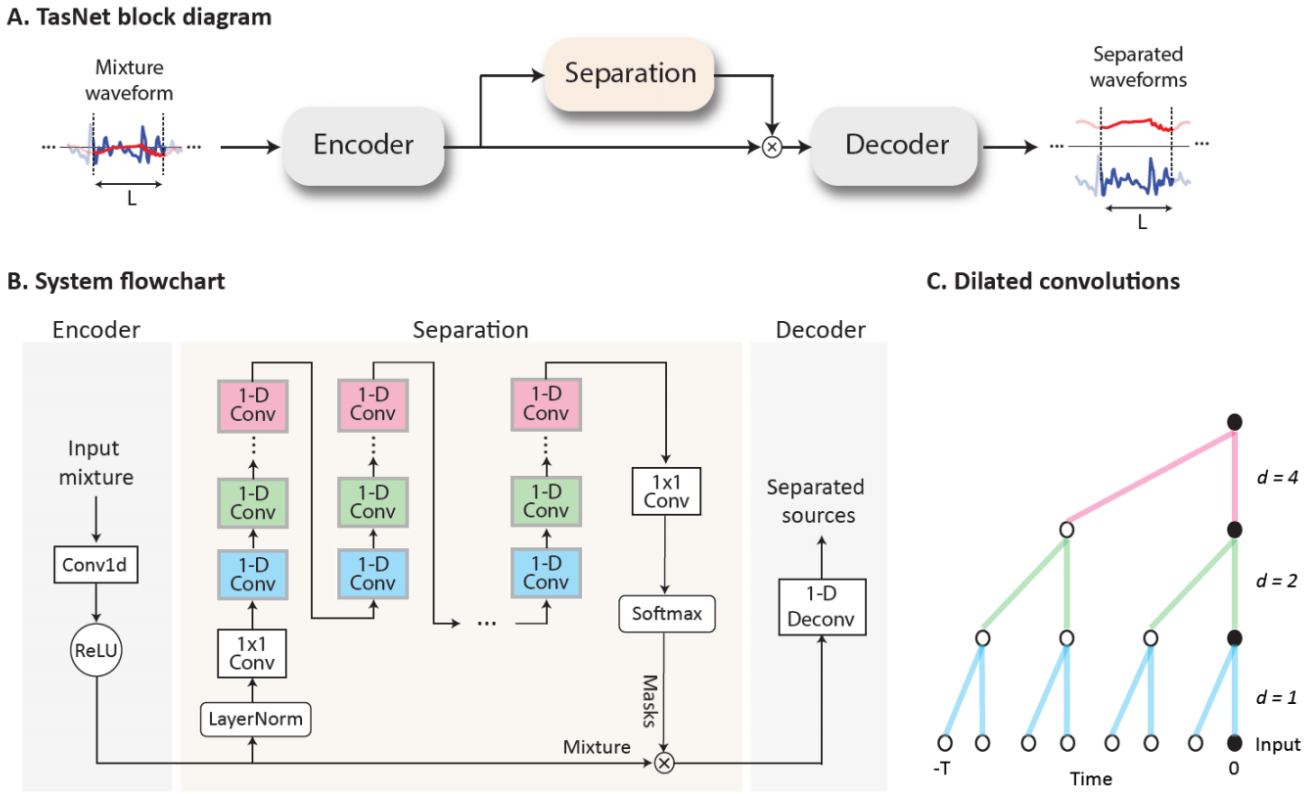
Speaker Extraction

- Target Speaker Extraction: only separate target speaker signal using reference audio from target speaker.





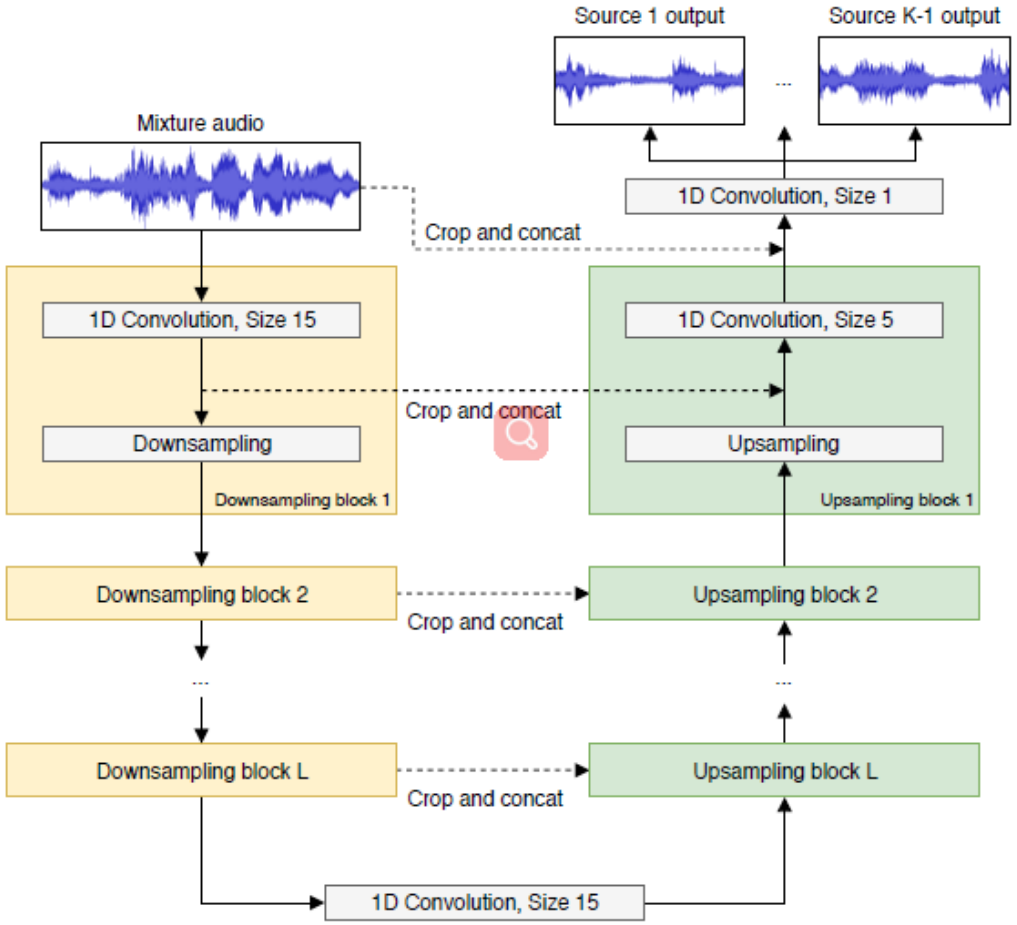
TasNet



- The drawbacks of time-frequency representation:
  - The decoupling of the phase and magnitude of the signal
  - The suboptimality of spectrogram representations for speech separation
  - The long latency in calculating the spectrogram

Luo Y , Mesgarani N . TasNet: Surpassing Ideal Time-Frequency Masking for Speech Separation[J]. 2018.

WAVE-UNet



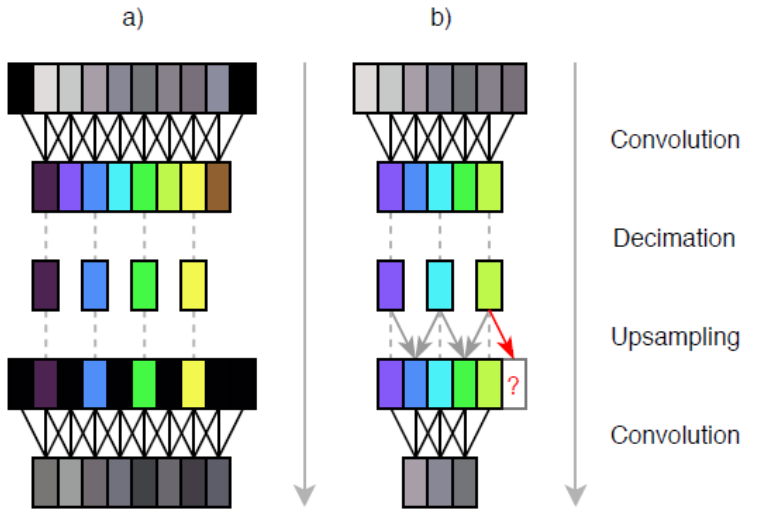
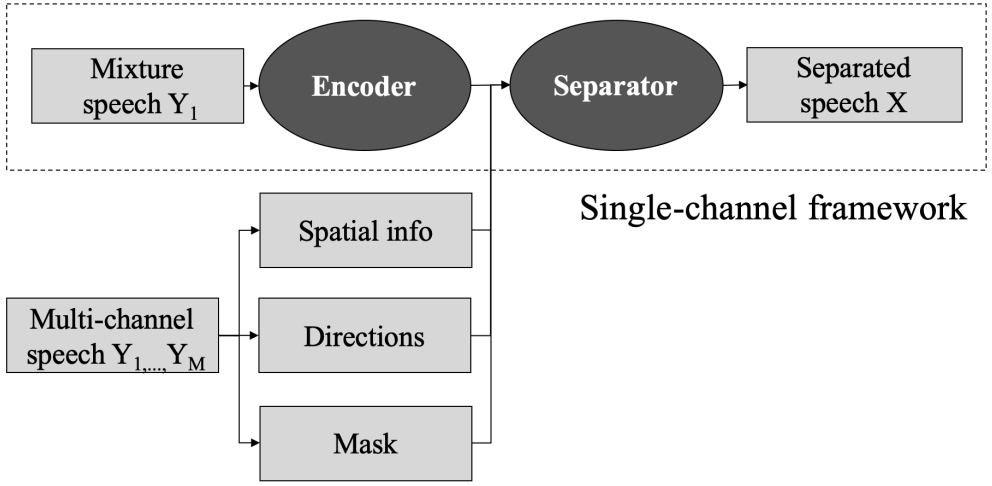


Figure 2. a) Common model (e.g. [7]) with an even number of inputs (grey) which are zero-padded (black) before convolving, creating artifacts at the borders (dark colours). After decimation, a transposed convolution with stride 2 is shown here as upsampling by zero-padding intermediate and border values followed by normal convolution, which likely creates high-frequency artifacts in the output. b) Our model with proper input context and linear interpolation for upsampling from Section 3.2.2 does not use zero-padding. The number of features is kept uneven, so that upsampling does not require extrapolating values (red arrow). Although the output is smaller, artifacts are avoided.

WAVE-U-NET: A MULTI-SCALE NEURAL NETWORK FOR END-TO-END AUDIO SOURCE SEPARATION

General Framework: Multi-channel

■ Separation based on Spatial Feature Extraction



Spectral (Single-channel) + Spatial (Multi-channel)

- Cross-Correlation Function (CCF) [2, 3]
- Interaural Time Difference (ITD) [2, 3, 13]
- Interaural Intensity Difference (IID) [13]
- Interaural Level Difference (ILD) [2, 3, 5]
- Interaural Phase Difference (IPD) [4, 5, 7-11, 14, 15], cosIPD [4, 8, 9], sinIPD [4, 8, 9], IPD variant [6, 12]
- Generalized Cross-Correlation (GCC) [1, 9]
- Angle features [10, 15, 16]
- Estimated Mask [5]

[1] A Study of Learning Based Beamforming Methods for Speech Recognition
[2] Binaural Classification for Reverberant Speech Segregation Using Deep Neural Networks
[3] Deep Learning Based Binaural Speech Separation in Reverberant Environments
[4] End-to-End Multi-Channel Speech Separation
[5] Exploring multi-channel features for denoising-autoencoder-based speech enhancement
[6] Iterative Deep Neural Networks for Speaker-Independent Binaural Blind Speech Separation
[7] Multi-band PIT and Model Integration for Improved Multi-channel Speech Separation
[8] Multi-Channel Block-Online Source Extraction Based on Utterance Adaptation
[9] Multi-Channel Deep Clustering: Discriminative Spectral and Spatial Embeddings for Speaker-Independent
[10] Multi-Channel Overlapped Speech Recognition with Location Guided Speech Extraction Network
[11] Multi-Microphone Neural Speech Separation for Far-Field Multi-Talker Speech Recognition
[12] Recognizing Overlapped Speech in Meetings: A Multichannel Separation Approach Using Neural Networks
[13] Speech Segregation based on Sound Localization
[14] Unsupervised Deep Clustering for Source Separation: Direct Learning from Mixtures Using Spatial Information
[15] Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information
[16] A Comprehensive Study of Speech Separation: Spectrogram vs Waveform Separation

Multi-channel SS

■ Separation based on Spatial Feature Extraction

➢ Cross-Correlation Function (CCF)

$$CCF(c,m,\tau)=\frac{\sum_k x_{cm,l}(k)x_{cm,r}(k-\tau)}{\sqrt{\sum_k x_{cm,l}^2(k)}\sqrt{\sum_k x_{cm,r}^2(k-\tau)}}$$

➢ Interaural Time Difference (ITD)

$$ITD(c,m)=\begin{pmatrix} CCF(c,m,\tilde{\tau}) \\ \max_{\tau} CCF(c,m,\tau) \end{pmatrix}$$

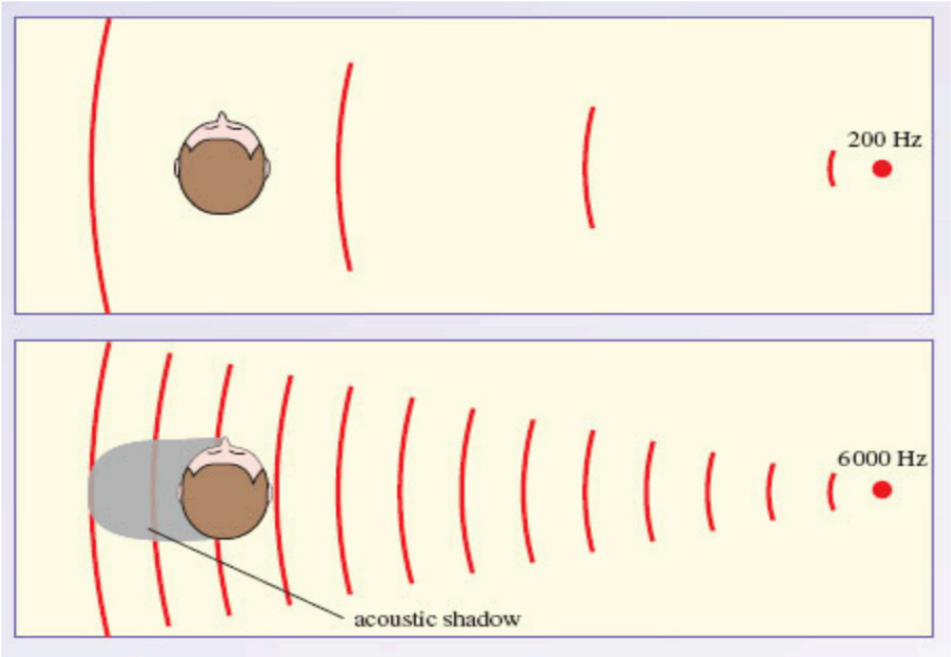
➢ Interaural Intensity Difference (IID)

➢ Interaural Level Difference (ILD)

$$ILD(c,m)=10\log_{10}\frac{\sum_k x_{cm,l}^2(k)}{\sum_k x_{cm,r}^2(k)}$$

PS: ITD are most effective at low frequency (<1.5kHz) and IID dominate the high frequency range.

Link: https://isle.hanover.edu/Ch11AudBrainLoc/Ch11InterLoud.html



In fact, interaural differences in intensity are negligible at low frequencies, but may be as large as 20 dB at high frequencies (**Figure).** Figure 45 Low-frequency tones are not affected by the listener's head, so the intensity of a 200 Hz tone is the same at both ears. High-frequency tones (e.g. 6000 Hz) are affected by the presence of the listener's head and result in an acoustic shadow that decreases the intensity of the tone reaching the listener's far ear.

Multi-channel SS

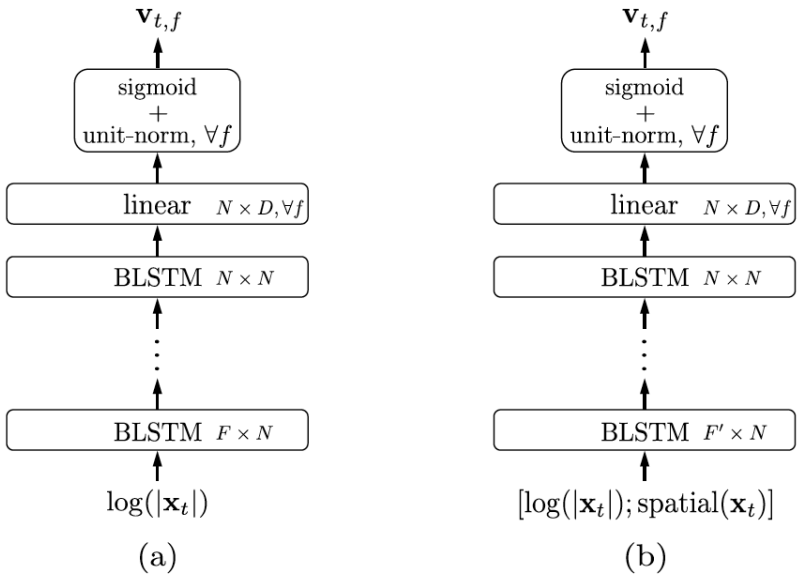


Fig. 1. Network architecture of (a) single-channel deep clustering, (b) multi-channel deep clustering.

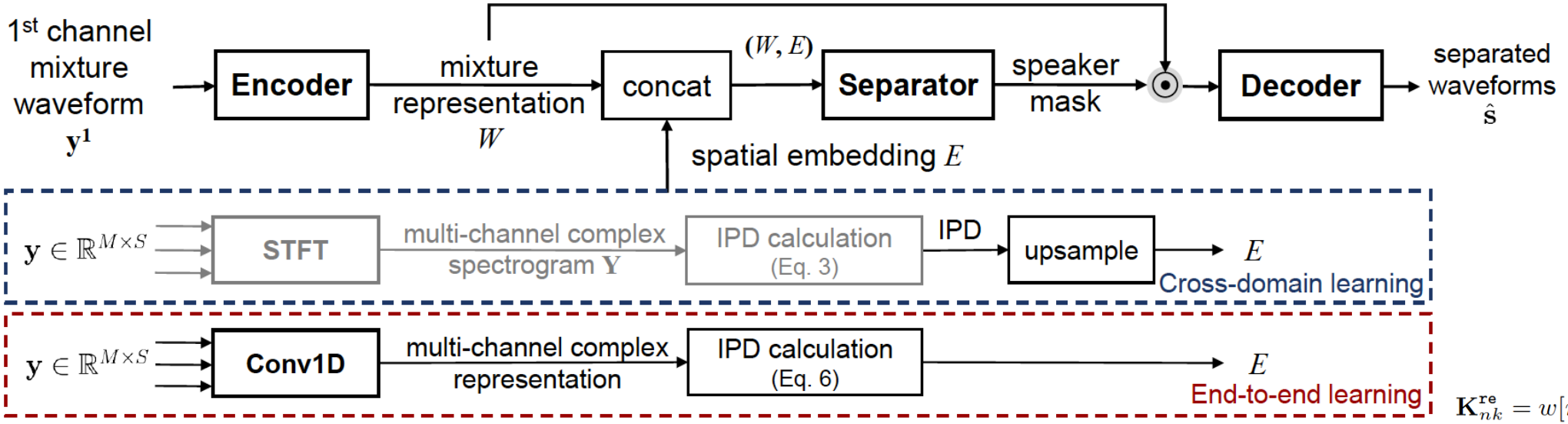
Table 1. SDR (dB) results on spatialized anechoic wsj0-2mix data

Approaches	Features	SDR
1ch Deep Clustering	Log mag.	10.3
2ch Deep Clustering	Log mag. + cosIPD	12.5
2ch Deep Clustering	Log mag. + cosIPD + sinIPD	12.9
2ch Deep Clustering	Log mag. + GCC	12.9
IRM/IBM	-	12.7/13.5

Table 2. SDR (dB) results on spatialized reverberant wsj0-2mix data

Approaches	Features	SDR
1ch Deep Clustering	Log mag.	6.9
2ch Deep Clustering	Log mag., GCC	7.5
	+ variance normalization	8.8
2ch Deep Clustering	Log mag., cosIPD	8.6
	Log mag., cosIPD, sinIPD	8.9
3ch Deep Clustering	Log mag., cosIPD, sinIPD	9.3
4ch Deep Clustering	Log mag., cosIPD, sinIPD	9.4
Oracle MCWF (2ch)	-	4.9
Oracle MCWF (3ch)	-	7.0
Oracle MCWF (4ch)	-	8.3
Oracle MCWF (5ch)	-	9.2
Oracle MCWF (6ch)	-	9.9
Oracle MCWF (7ch)	-	10.5
Oracle MCWF (8ch)	-	10.9
MESSL [18]	see [18]	3.3
GCC-NMF [19]	see [19]	2.7
IRM/IBM	-	11.9/12.7

Cross-domain learning



- Cross-domain learning

$$y[n] \xrightarrow{\text{STFT}} \mathbf{Y}_{nk} = \sum_{m=0}^{T-1} y[m] w[n-m] e^{-i\frac{2\pi m}{T}k}$$

$$\text{IPD}_{nk}^{(u)} = \angle \mathbf{Y}_{nk}^{u_1} - \angle \mathbf{Y}_{nk}^{u_2}$$

- End-to-end learning

$$\mathbf{K}_{nk}^{\text{re}} = w[n]\cos(2\pi nk/T)$$
$$\mathbf{K}_{nk}^{\text{im}} = w[n]\sin(2\pi nk/T) \quad (5)$$

$$y[n] \xrightarrow{\text{STFT}} \mathbf{Y}_{nk} = \overbrace{e^{-i\frac{2\pi n}{T}k}}^{\text{phase factor}} \left(y[n] \circledast \overbrace{w[n]e^{i\frac{2\pi n}{T}k}}^{\text{STFT kernel}}\right) \quad (4)$$

$$\text{IPD}_{nk}^{(u)} = \arctan\left(\frac{y^{u_1} \circledast \mathbf{K}_{nk}^{\text{re}}}{y^{u_1} \circledast \mathbf{K}_{nk}^{\text{im}}}\right) - \arctan\left(\frac{y^{u_2} \circledast \mathbf{K}_{nk}^{\text{re}}}{y^{u_2} \circledast \mathbf{K}_{nk}^{\text{im}}}\right) \quad (6)$$

- PS: The selected pairs for IPDs are (1, 4), (2, 5), (3, 6), (1, 2), (3,4) and (5, 6) in all experiments

[1] End-to-End Multi-Channel Speech Separation

Multi-channel SS

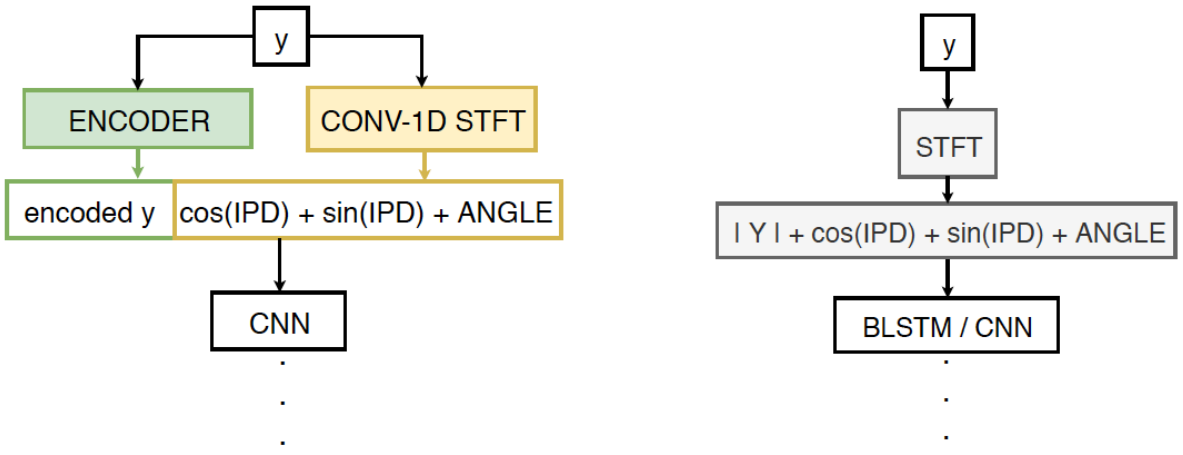


Figure 1: *Multi-channel speech separation; (left) waveform separation, (right) spectrogram separation.*

- (Example) Motivation for TGT: Using directional information or speaker-related knowledge can help to extract a target speaker
- IPD features:

$$\text{IPD}_{i,t,f} = \angle\left(\frac{Y_{i_1,t,f}}{Y_{i_2,t,f}}\right), i = 1:6$$

- Angle features:

$$A_{s,t,f} = \sum_{i=1}^{6} \frac{e_s^{i,f} \frac{Y_{i_1,t,f}}{Y_{i_2,t,f}}}{\left| e_s^{i,f} \frac{Y_{i_1,t,f}}{Y_{i_2,t,f}} \right|}, \quad (4)$$

where s is the speaker index and $e_s^{i,f}$ represents steering vector coefficient of speaker s direction of arrival at microphone i for frequency f [3].

[2] A Comprehensive Study of Speech Separation: Spectrogram vs Waveform Separation

Multi-channel SS

Table 1: *Experimental setup for time/frequency-domain models.*

Model	Input	# of Parameters	Setting	Normalization
F-CNN-1	$\|Y_0\|$	8.78M	N=257	gLN
F-CNN-2	$\|Y_0\|$ + cos(IPD) + sin(IPD)	9.58M	N=257×13	gLN
F-CNN-3	$\|Y_0\|$ + cos(IPD) + sin(IPD) + Angle	9.71M	N=257×15	gLN
F-BLSTM-1	$\|Y_0\|$	67.05M	4 × BLSTM-896	-
F-BLSTM-2	$\|Y_0\|$ + cos(IPD) + sin(IPD)	89.15M	4 × BLSTM-896	-
F-BLSTM-3	$\|Y_0\|$ + cos(IPD) + sin(IPD) + Angle	92.84M	4 × BLSTM-896	-
T-CNN-1	y_0	8.76M	L/N=40/256	gLN
T-CNN-2	y_0 + cos(IPD) + sin(IPD)	8.83M	L/N=40/256	BN

Table 2: *Comparing CNN separation network and Si-SNR loss function in spectrogram separation with single-channel input.*

Model	Loss	Si-SNR					SDR				
		0-15°	15-45°	45-90°	90-180°	AVG	0-15°	15-45°	45-90°	90-180°	AVG
F-BLSTM-1	uPIT-SiSNR	7.54	7.80	7.72	7.81	7.74	8.14	8.39	8.29	8.38	8.32
	uPIT-MSE	6.50	6.79	6.67	6.78	6.71	6.98	7.24	7.11	7.22	7.16
F-CNN-1	uPIT-SiSNR	7.08	7.48	7.45	7.48	7.42	7.70	8.06	8.02	8.06	8
	uPIT-MSE	6.31	6.67	6.54	6.65	6.58	6.81	7.13	7.0	7.1	7.04

[2] A Comprehensive Study of Speech Separation: Spectrogram vs Waveform Separation

Multi-channel SS

Table 3: $|Y_0|$+*cos(IPD)+sin(IPD)+Angle are input to the network to evaluate the performance of multi-channel framework in extracting the speaker of interest.*

Model	Loss	Si-SNR					SDR				
		0-15°	15-45°	45-90°	90-180°	AVG	0-15°	15-45°	45-90°	90-180°	AVG
F-BLSTM-3	uPIT-SiSNR	5.09	9.11	9.62	9.31	8.72	5.81	9.64	10.11	9.85	9.27
	TGT-SiSNR	4.59	9.07	9.59	9.13	8.58	5.31	9.60	10.08	9.64	9.12
F-CNN-3	uPIT-SiSNR	5.59	9.38	10.29	10.79	9.49	6.92	10.36	11.2	11.75	10.50
	TGT-SiSNR	4.88	9.69	10.32	9.80	9.2	5.61	10.19	10.77	10.27	9.71

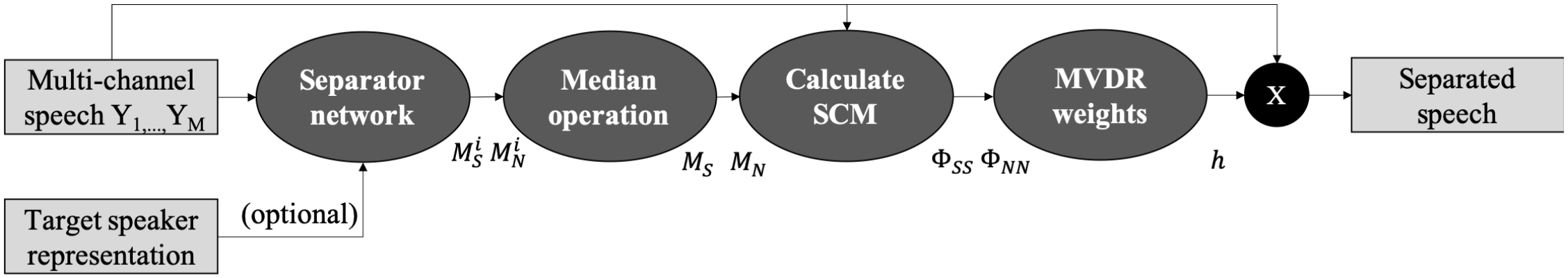
Table 4: *Comparing spectrogram and waveform separation for both separation and ASR tasks.*

# Channels	Domain	Model	Si-SNR					SDR					PESQ	WER Reduc. (%)
			0-15°	15-45°	45-90°	90-180°	AVG	0-15°	15-45°	45-90°	90-180°	AVG		
1-ch	time	T-BLSTM	-	-	-	-	-	-	-	-	-	-	-	-
		T-CNN-1	9.02	9.33	9.59	9.71	**9.47**	9.57	9.83	10.09	10.2	**9.97**	1.95	45.53
	freq	F-BLSTM-1	7.54	7.80	7.72	7.81	**7.74**	8.14	8.39	8.29	8.38	**8.32**	1.77	32.21
		F-CNN-1	7.08	7.48	7.45	7.48	**7.42**	7.70	8.06	8.02	8.06	**8**	1.77	35.17
m-ch	time	T-BLSTM	-	-	-	-	-	-	-	-	-	-	-	-
		T-CNN-2	7.70	11.63	12.33	12.62	**11.55**	8.31	12.07	12.74	13.03	**11.99**	2.10	59.11
	freq	F-BLSTM-2	5.41	9.37	10.13	10.65	**9.38**	6.13	9.89	10.62	11.13	**9.91**	1.92	45.32
		F-CNN-2	6.88	10.27	11.02	11.54	**10.36**	7.5	10.75	11.47	11.99	**10.84**	2.00	51.73
Oracle IBM			11.56	11.51	11.53	11.53	**11.53**	11.93	11.86	11.88	11.88	**11.88**	2.01	50.37
Oracle IAM			11.05	11.03	11.05	11.03	**11.04**	11.33	11.3	11.31	11.29	**11.30**	2.23	71.45
Oracle IRM			11.01	10.96	10.98	10.97	**10.98**	11.45	11.39	11.39	11.39	**11.40**	2.22	70.28
Oracle IPSM			13.68	13.6	13.64	13.63	**13.63**	14.04	13.94	13.98	13.97	**13.98**	2.28	71.10
Reference													2.35	73.01

[2] A Comprehensive Study of Speech Separation: Spectrogram vs Waveform Separation

General Framework: Multi-channel

- Time-frequency masking for beamforming



Multi-channel SS

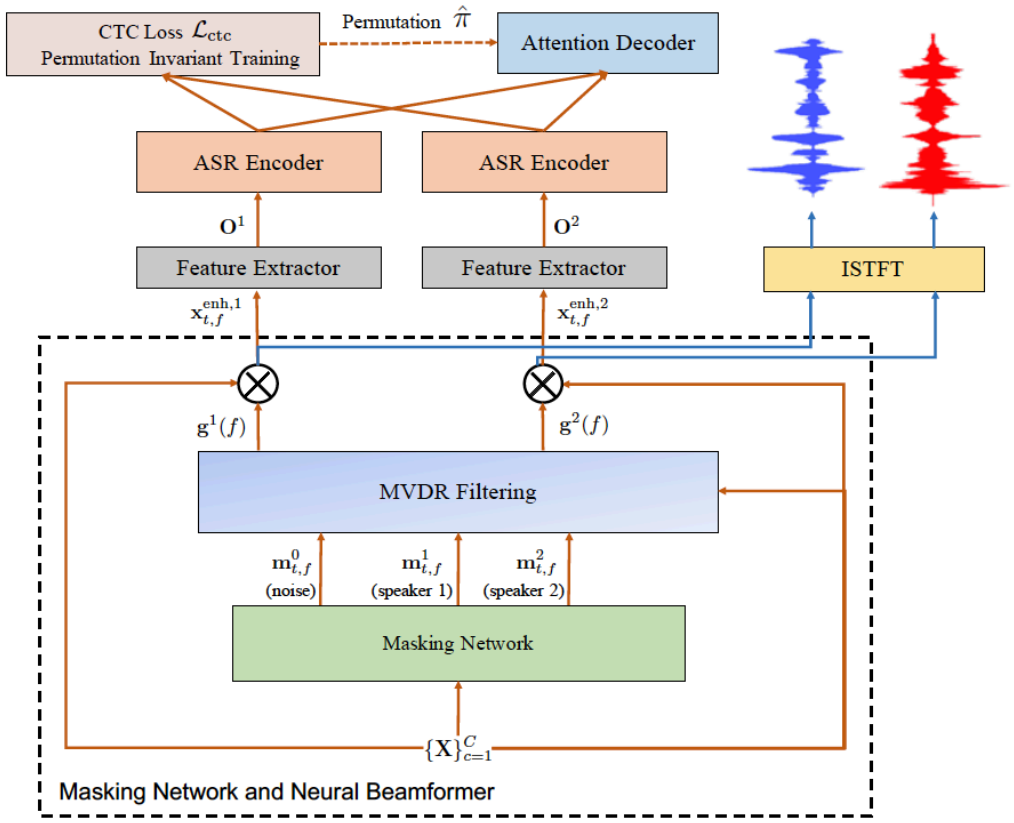


Fig. 1. End-to-End Multi-channel Multi-speaker Model

- Multi-source Neural Beamformer

$$\mathbf{\Phi}^i(f) = \frac{1}{\sum_{t=1}^T \mathbf{m}_{t,f}^i} \sum_{t=1}^T \mathbf{m}_{t,f}^i \mathbf{x}_{t,f} \mathbf{x}_{t,f}^H \in \mathbb{C}^{C\times C},$$

$$\mathbf{g}^i(f) = \frac{(\sum_{j\neq i} \mathbf{\Phi}^j(f))^{-1}\mathbf{\Phi}^i(f)}{\text{Tr}((\sum_{j\neq i} \mathbf{\Phi}^j(f))^{-1}\mathbf{\Phi}^i(f))}\mathbf{u} \in \mathbb{C}^C,$$

$$\hat{s}_{t,f}^i = (\mathbf{g}^i(f))^H \mathbf{x}_{t,f} \in \mathbb{C}.$$

- E2E speech recognition

$$\mathcal{L} = \lambda\mathcal{L}_{\text{ctc}} + (1-\lambda)\mathcal{L}_{\text{att}},$$
$$\mathcal{L}_{\text{ctc}} = \sum_i \text{Loss}_{\text{ctc}}(\mathbf{Z}^i, \mathbf{R}^{\hat{\pi}(i)}),$$
$$\mathcal{L}_{\text{att}} = \sum_i \text{Loss}_{\text{att}}(\mathbf{Y}^i, \mathbf{R}^{\hat{\pi}(i)}),$$

Summary and Discussion

- **Challenges: speech rarely occurs in isolation in real cases.**

  Solutions:

  - **Speech separation framework + Human perception mechanism**
  - **Single-channel framework + Multi-channel stucture**
  - **Frequency-domain + Time-domain**

- **Future direction: focus on real cases**

  - **Task-driven speech separation framework**
  - **Frequency-domain → Time-domain**
  - **Multi-view strategy**


Thank You!