

Selective Auditory Attention: Speech Separation and Speaker Extraction

Haizhou Li

National University of Singapore, Singapore

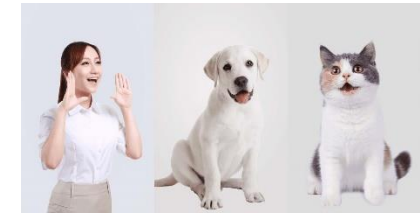


Agenda

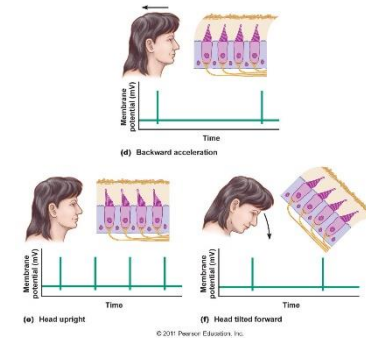
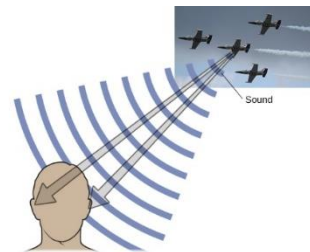
- **Selective auditory attention**
- **Speech separation & speaker extraction**
- **Applications**

Ears and Hearing

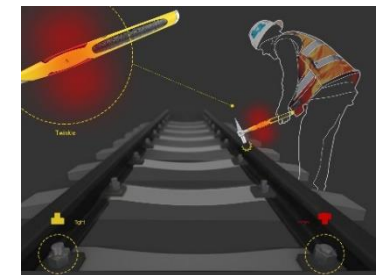
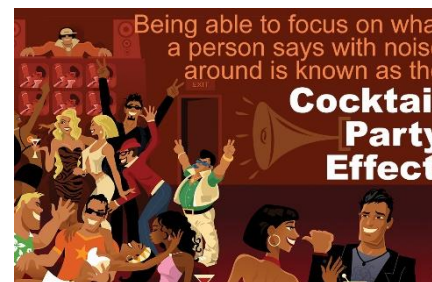
Frequency Analyser



Localization/Equilibrium

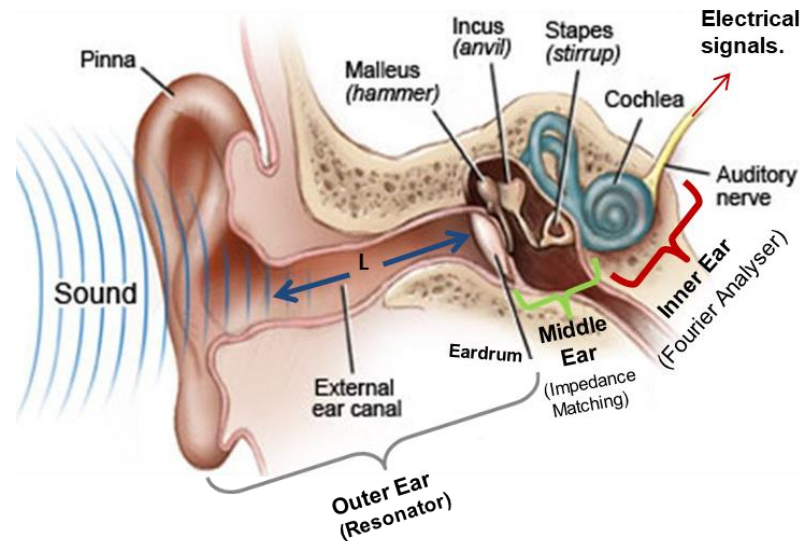


Attention



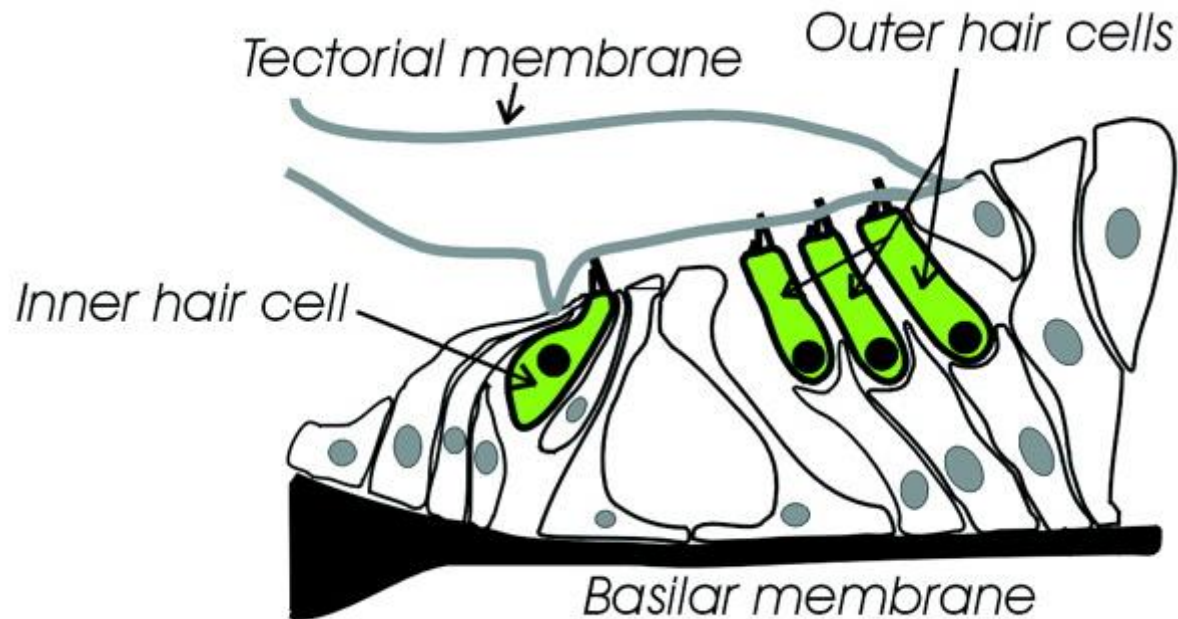
From the Perspective of Physiology

Human Auditory System



- ✓ The human outer ear is most sensitive at about 3kHz and provides about 20dB (decibels) of gain to the eardrum at around 3kHz.
- ✓ Middle ear transforms the vibrating motion of the eardrum into motion of the stapes via the two tiny bones, the malleus and incus .
- ✓ The combined frequency response of the outer and middle ear is a band-pass response, with its peak dominated near 3 kHz.

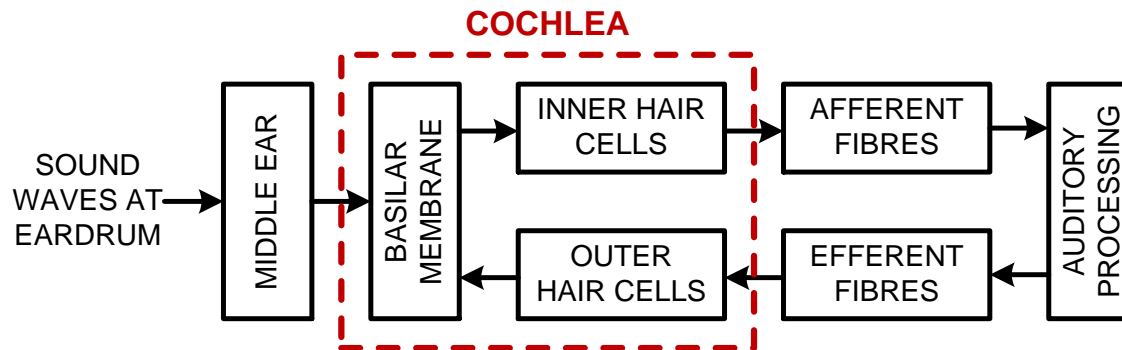
Structure of the Cochlea



- 3,500 inner hair cells (30 dB), 12,000 outer hair cells
- 14 Micro Watt

L. Trussell, Mutant ion channel in cochlear hair cells causes deafness, PNAS Apr. 11, 2000 97 (8) 3786-3788;

Auditory Sensors and Actuators



- Both passive and active systems
- The outer hair cells (OHC) provide this active mechanism - they amplify the motion picked up by the IHC
- Dynamic range 120 dB/0.5dB
- Frequency range 10 octaves/0.3 octave

Passive and Active Pathways

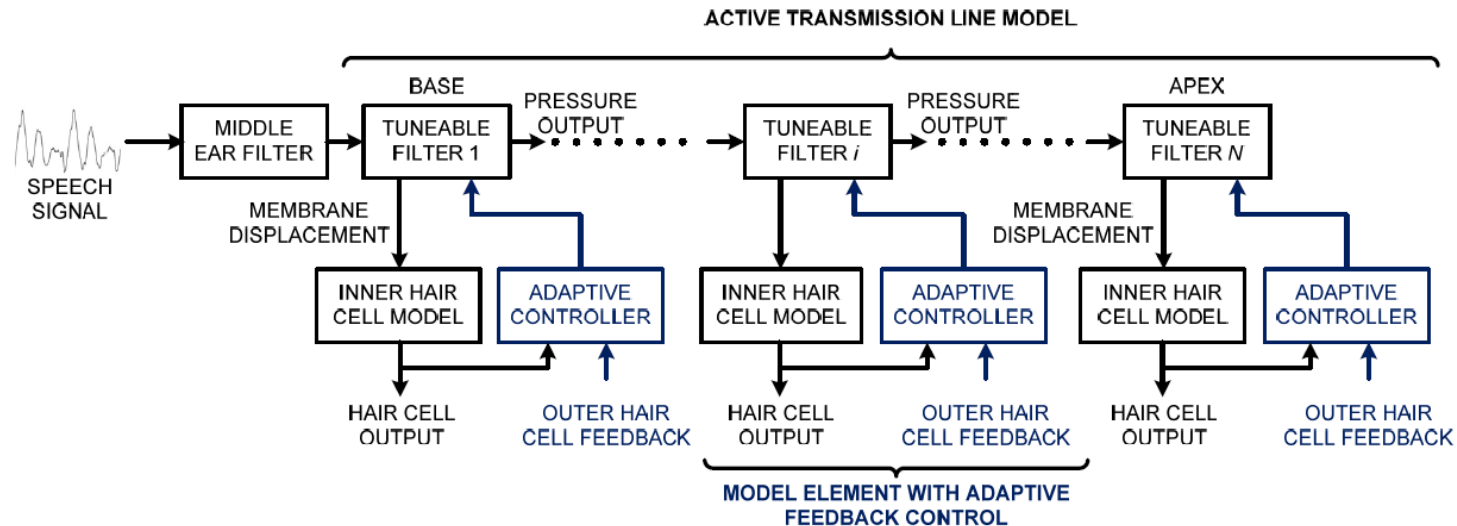


Figure 2: Transmission line model as it appeared in [39], showing sound intensity-based feedback loop and adaptive gain to sharpen the tuning curves in response to the input. Despite its many advantages, it has never been employed in speech processing systems.

E. Ambikairajah and B. McDonagh, "An Active Model of the Auditory Periphery with Realistic Temporal and Spectral Characteristics," in *5th Australian International Conference on Speech Science and Technology*, 1994.

Trainable Feature Frontends

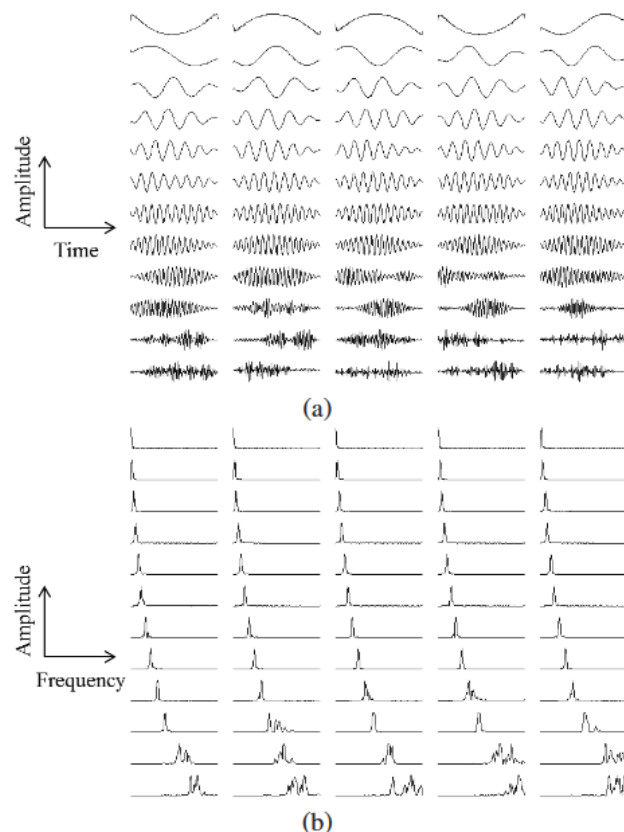


Figure 3: Examples of the subband filters trained on the ESC-50 database: (a) subband filters in the time-domain (i.e., impulse responses), (b) subband filters in the frequency-domain.

N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," INTERSPEECH, 2015

H. B. Sailor and H. A. Patil, "Novel unsupervised auditory filterbank learning using convolutional RBM for speech recognition," IEEE/ACM T-ASLP, vol. 24, no. 12, pp. 2341-2353, 2016

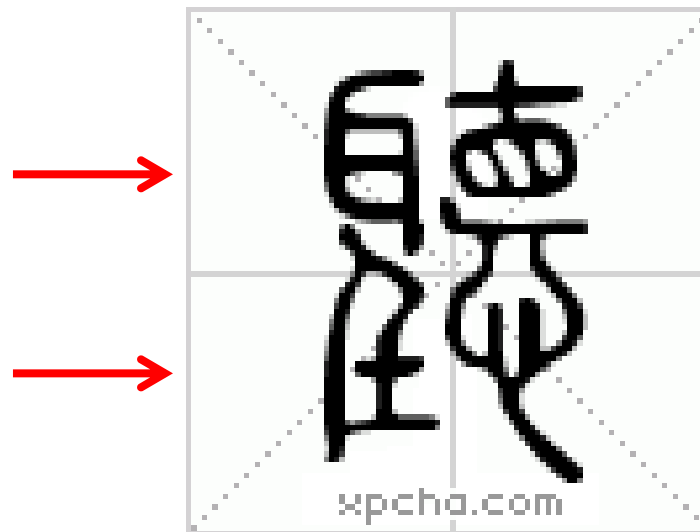
Neil Zeghidour et al, Learning Filterbanks from Raw Speech for Phone Recognition, ICASSP 2018

From the Perspective of Neuroscience

Chinese Character 'to listen'

Feature
Extraction
and Acoustic
Modeling
(the ear)

A person

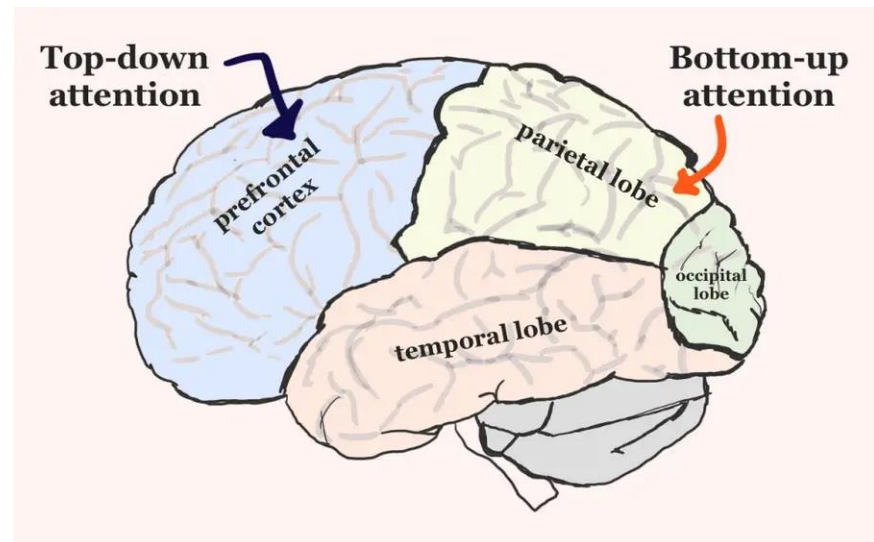


Undivided
attention
(the eyes)

Language
modeling
(the heart)

Control of Attention

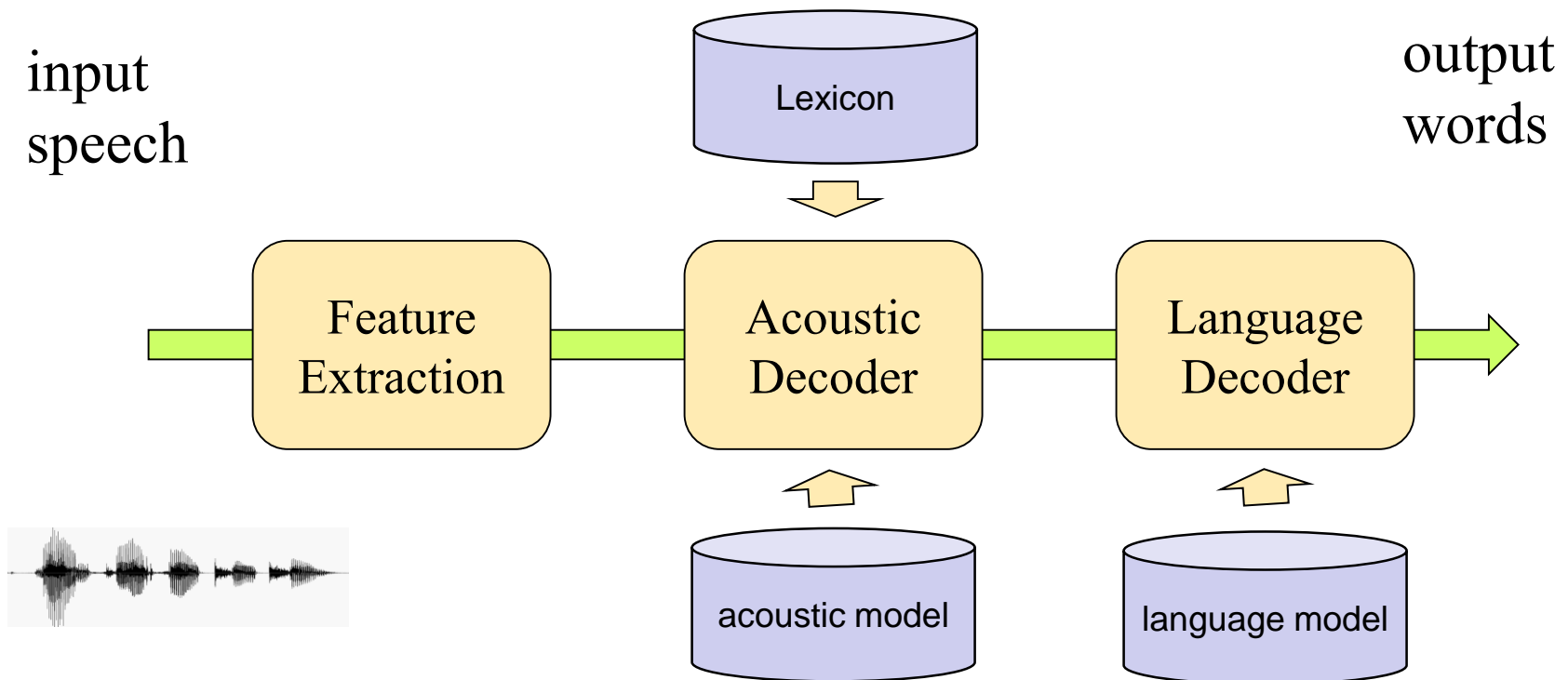
- **Top-down (or 'voluntary focus')**
- **Bottom-up (or 'stimulus-driven focus')**
- **Modulation by voluntary focus through Spectro-Temporal Receptive Fields**




Timothy J. Buschman, Earl K. Miller, Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices, *Science* 315(5820):1860-2 · March 2007

Speech Recognition Pipeline

- A view of static processing in sensory cortex with 'stimulus-driven focus'



From the Perspective of Psychoacoustics



Being able to focus on what
a person says with noise
around is known as the

Cocktail Party Effect.

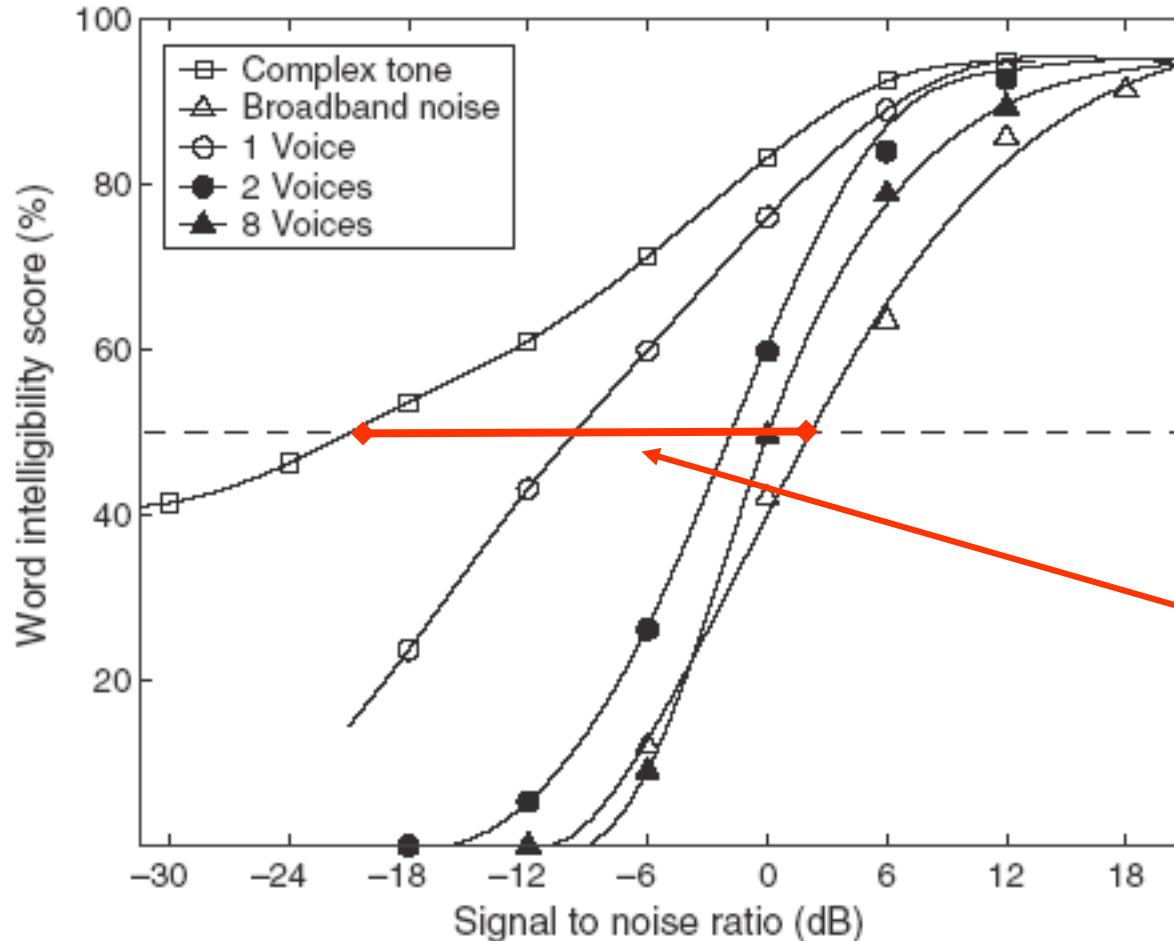
Source: <https://psychologenie.com/overview-of-cocktail-party-effect-in-psychology>

Cocktail Party

- “One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. This is such a common experience that we may take it for granted; we may call it ‘the cocktail party problem.’ **No machine has been constructed to do just that.**” (Cherry, 1957)
- “For ‘cocktail party’-like situations... when all voices are equally loud, speech remains intelligible for normal-hearing listeners even when there are as many as *six* interfering talkers” (Bronkhorst & Plomp’92)

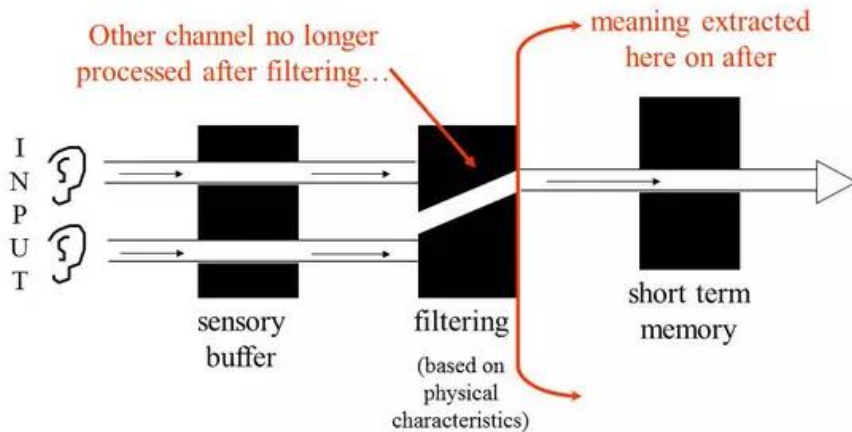


Effects of Competing Source

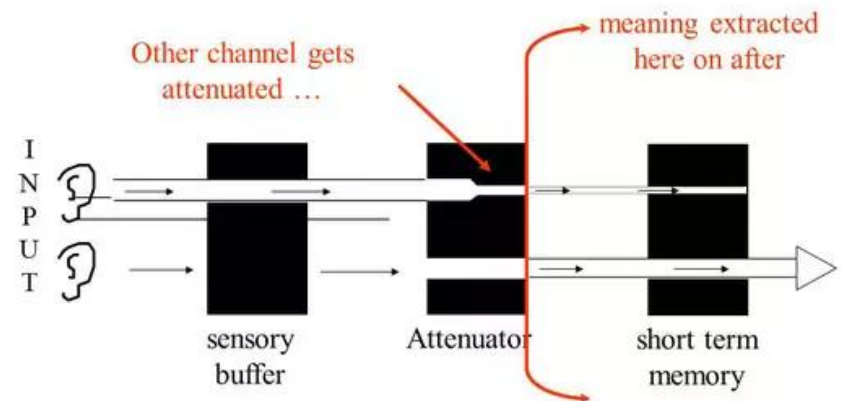


Speech
Reception
Threshold
Difference
(23 dB!)

Theory of Filtering



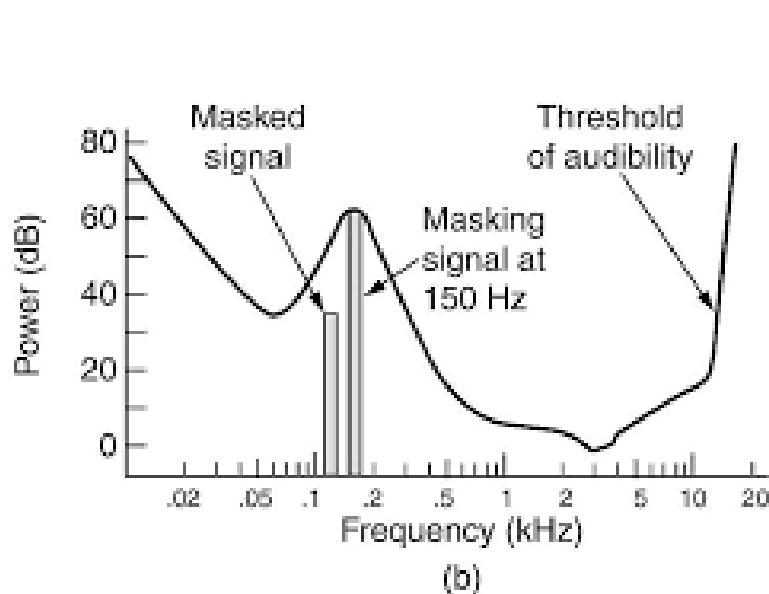
Broadbent (1958)



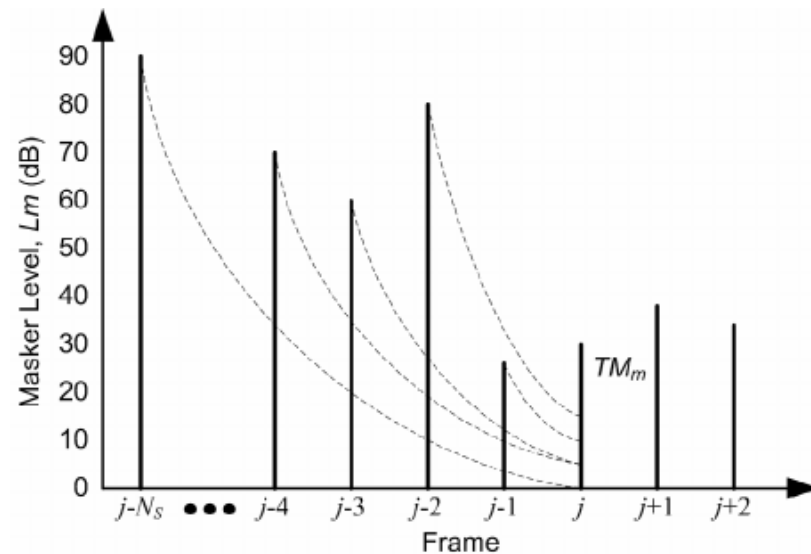
Treisman (1964)

Auditory Masking

- With the understanding of auditory mask, we can retain parts of a target sound that are stronger than the acoustic background, and discard the rest



Simultaneous Masking



Temporal Masking

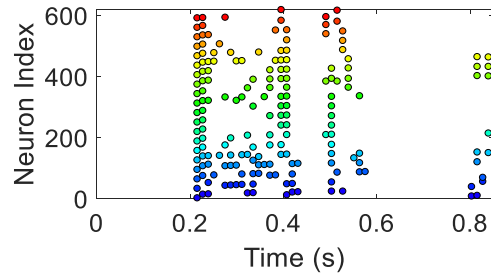
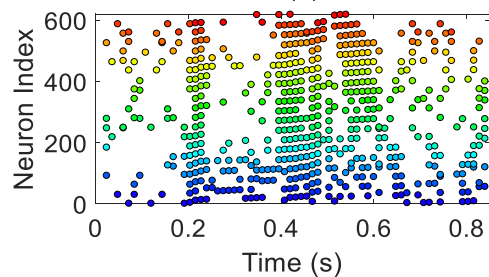
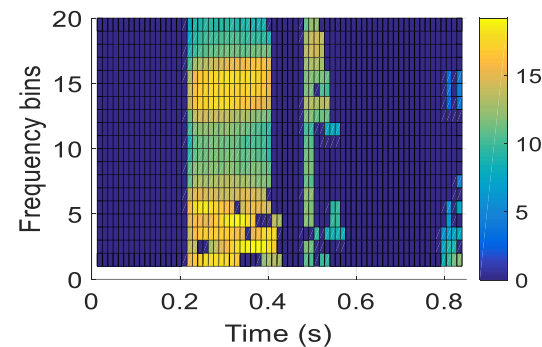
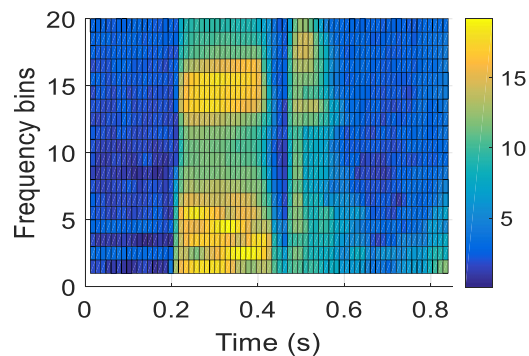
From the Perspective of Signal Processing

Auditory Masking

- **Example (not the best): Ideal Binary Mask (IBM)**
 - $s(t, f)$: Target energy in unit (t, f)
 - $n(t, f)$: Noise energy
 - θ : A local SNR criterion (LC) in dB, typically chosen to be 0 dB
 - It does not actually separate the mixture!

$$IBM(t, f) = \begin{cases} 1 & \text{if } s(t, f) - n(t, f) \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

Masking Effect



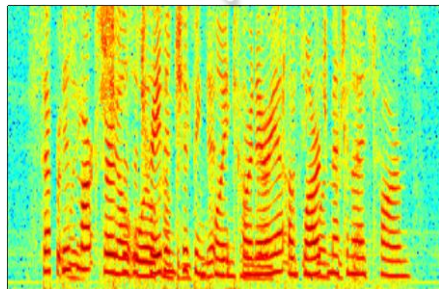
Original



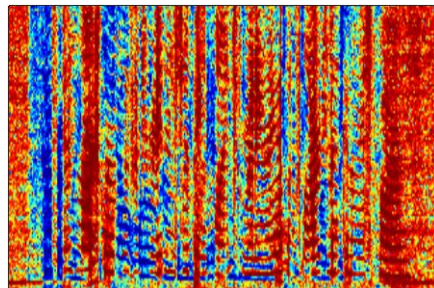
After Masking

| Dataset | RWCP | TIDIGITS | TIMIT |
|------------------|--------|----------|--------|
| Energy reduction | 39.38% | 50.48% | 29.33% |

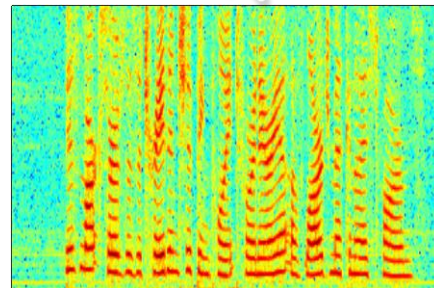
Masking Effect (Ideal Ratio Mask)



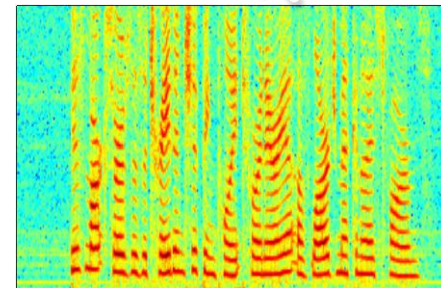
Female-Female Mixed Speech



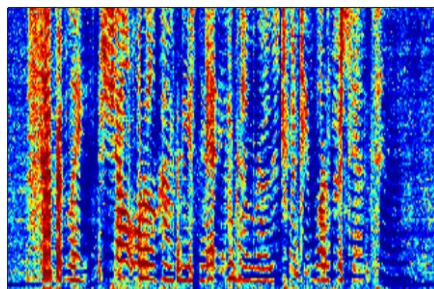
IRM for Speaker 1



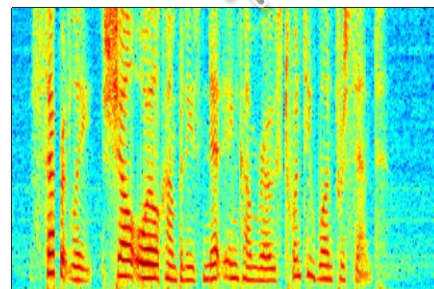
Separated Speech of Speaker 1 with IRM



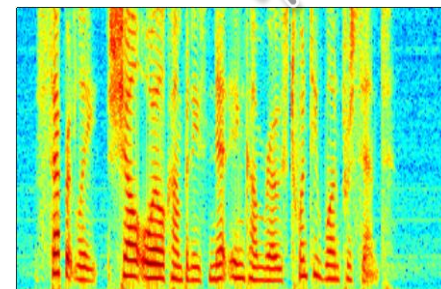
Clean Speech of Speaker 1



IRM for Speaker 2

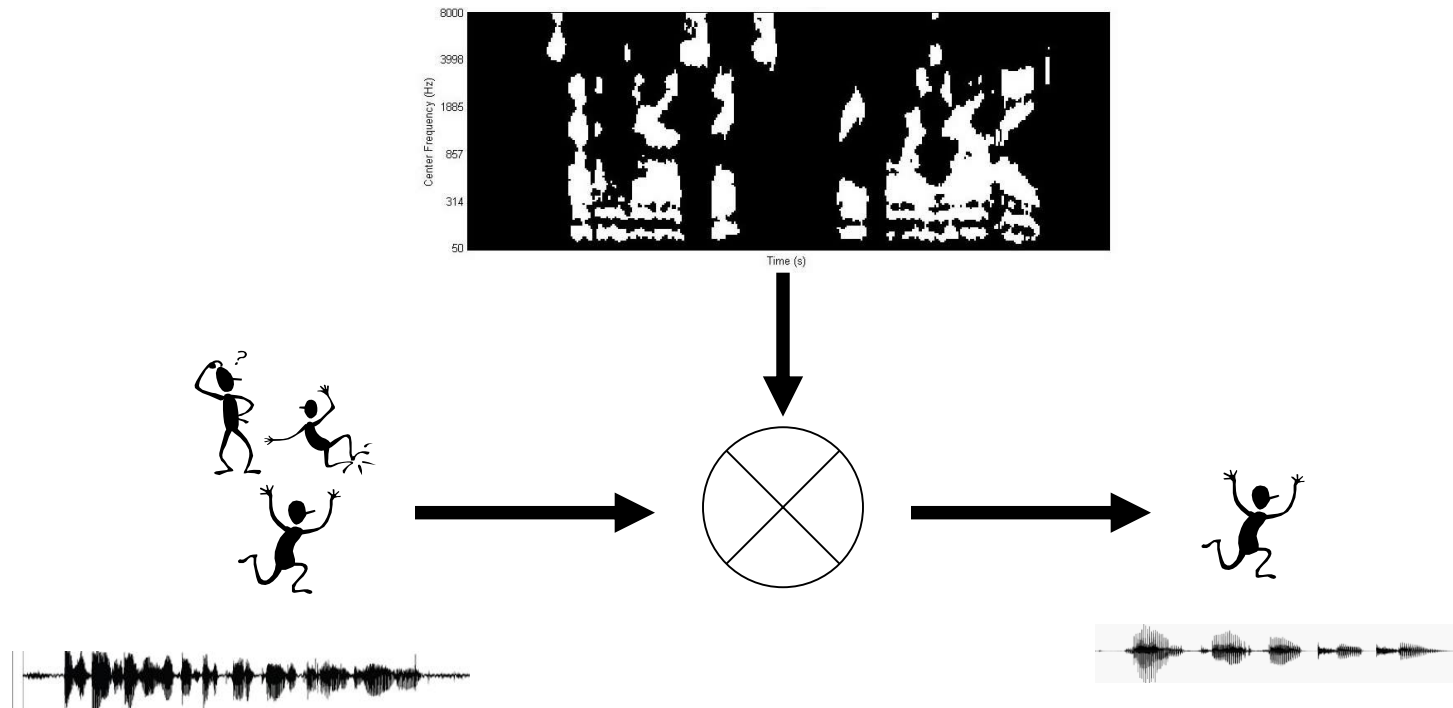


Separated Speech of Speaker 2 with IRM



Clean Speech of Speaker 2

Question: How to find the mask?

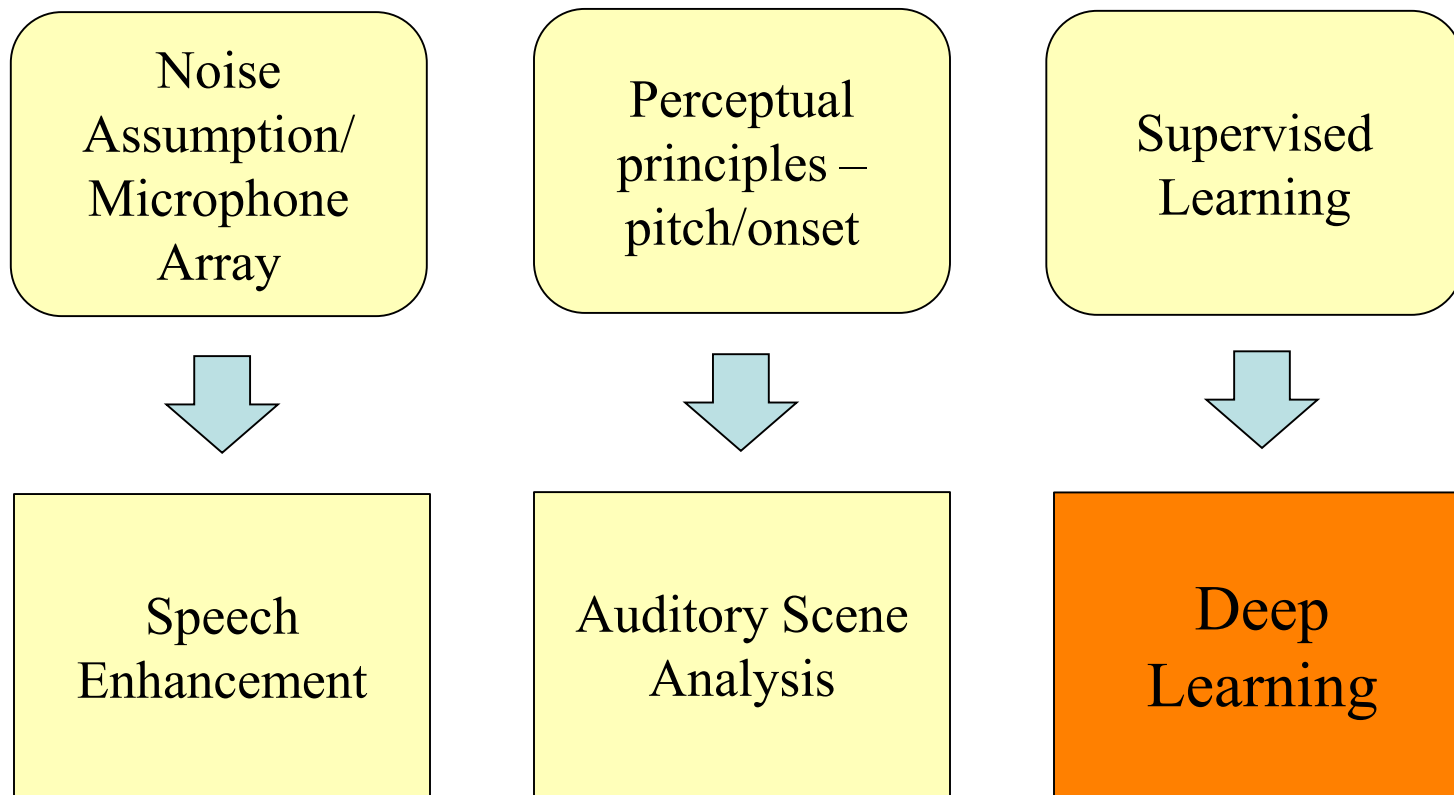


Spectro-Temporal Receptive Fields that reflect the temporal and spectral modulations belonging to the target sound events

Agenda

- Selective auditory attention
- **Speech separation & speaker extraction**
- Applications

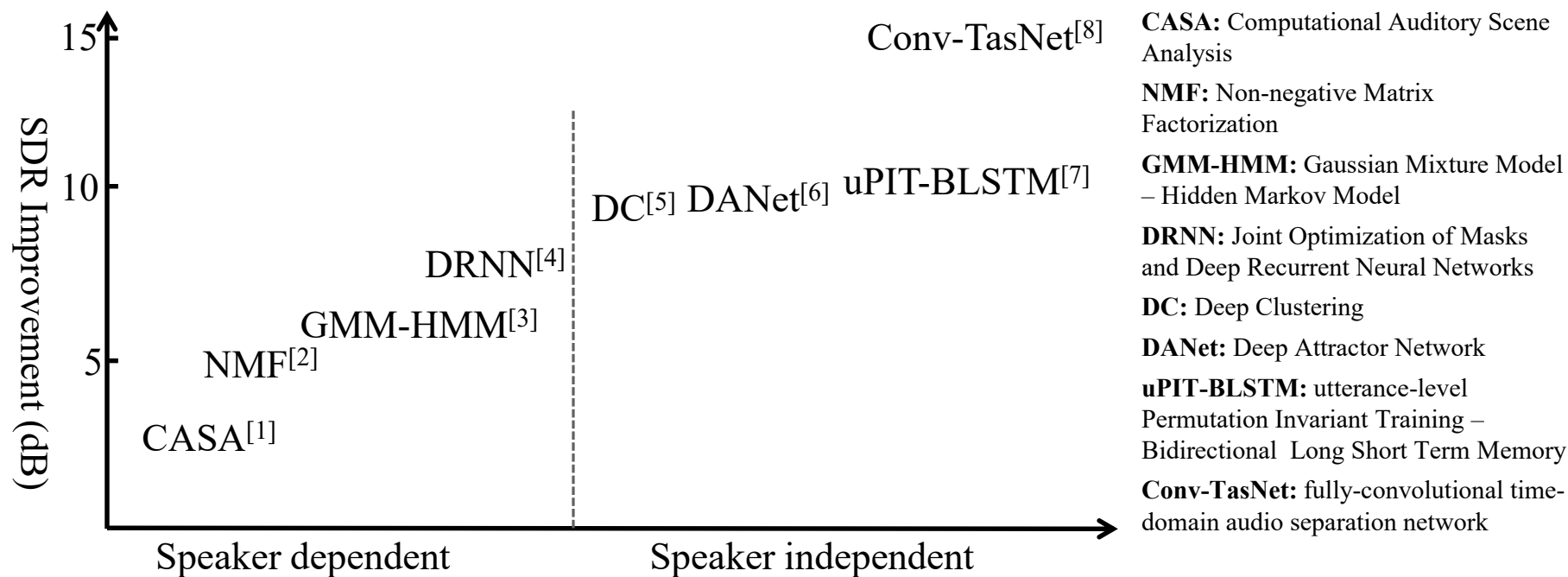
Monaural Speech Separation: Overview



[1] Albert S. Bregman, Auditory Scene Analysis. Cambridge, MA, USA: MIT Press, 1990.

[2] D. Wang, Supervised Speech Separation Based on Deep Learning: An Overview, IEEE/ACM T-ASLP 2018

Monaural Speech Separation: Overview



[1] D. P. Eills, “Prediction-driven computational auditory scene analysis”, PhD. Dissertation, MIT, 1996

[2] M. N. Schmidt and R. K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization”, in *Proc. INTERSPEECH 2006*

[3] T. Virtanen, “Speech recognition using factorial hidden markov models for separation in the feature space”, in *Proc. INTERSPEECH 2006*

[4] P.-S. Huang, M. Kim, M. Hasengawa-Johnson and P. Smaragdis, “Joint optimization of masks and deep recurrent neural networks for monaural source separation”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.23, No.12, 2015

[5] J. R. Hershey, et al, “Deep clustering: Discriminative embeddings for segmentation and separation”, in *Proc. ICASSP*, 2016, pp. 31-35

[6] Z. Chen, Y. Luo and N. Mesgarani, “Deep attractor network for single microphone speaker separation”, in *Proc. ICASSP*, 2017

[7] M. Kolbek, Dong Yu, Z.-H. Tan and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.25, No.10, pp.1901-1913, 2017

[8] Y. Luo and N. Mesgarani, “Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.27, No.8, pp.1256-1266, 2019

Idea 1: Deep Clustering

- DC^[1]: to project the spectrogram of the mixture to an embedding space, where the T-F bins belonging to the same speaker are grouped together.
- Objective function:
$$C_Y(V) = \|VV^T - YY^T\|_F^2 = \sum_{i,j} (\langle v_i, v_j \rangle - \langle y_i, y_j \rangle)^2$$
- To optimize towards the **ideal binary mask**

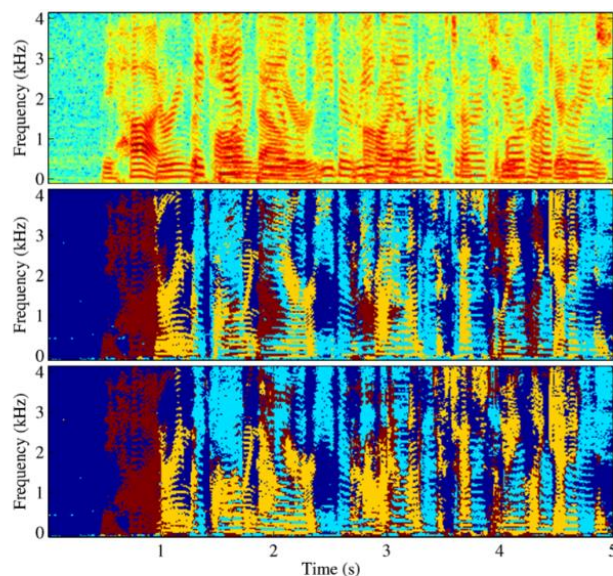
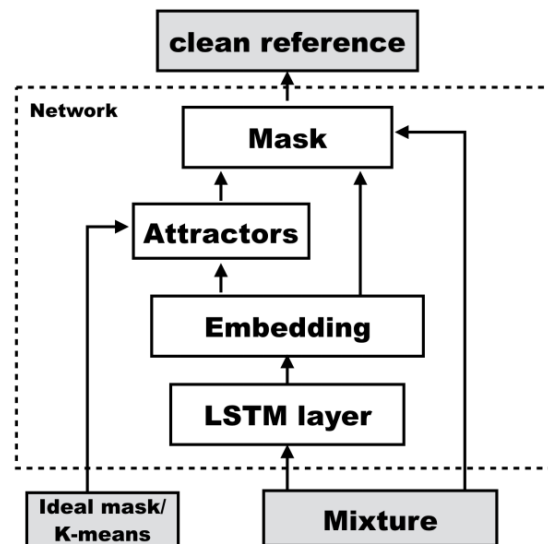


Figure 1: An example of three-speaker separation. Top: log spectrogram of the input mixture. Middle: ideal binary mask for three speakers. The dark blue shows the silence part of the mixture. Bottom: output mask from the proposed system trained on two-speaker mixtures.

[1] J. R. Hershey, Z. Chen, J. L. Roux and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation", in *Proc. ICASSP*, 2016, pp. 31-35

Idea 2: Deep Attractor Net

- DANet^[1]: Create attractor points in high dimensional embedding space and estimate masks within the network to separate signals directly.
- To implement the idea of **perceptual magnet effect**
- To minimize the **signal reconstruction error**

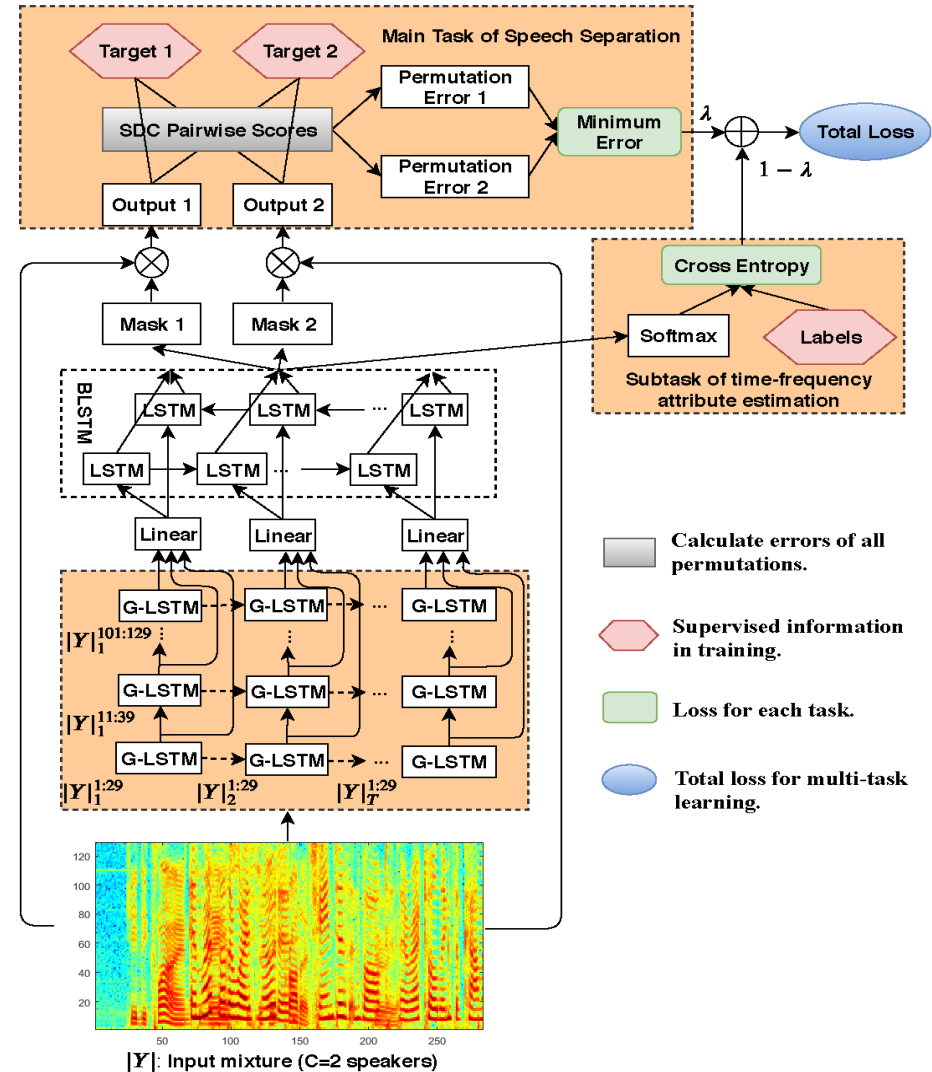


[1] Z. Chen, Y. Luo and N. Mesgarani, "Deep attractor network for single microphone speaker separation", in Proc. ICASSP, 2017

[2] Patricia K. Kuhl, "Human adults and human infants show a perceptual magnet effect," Perception & psychophysics, 50.2(1991): 93-107

Idea 3: Permutation-Invariant Training

- To estimate the mask by optimizing the direct signal reconstruction error and towards the actual mask
- Permutation invariant training to address speaker permutation problem during training



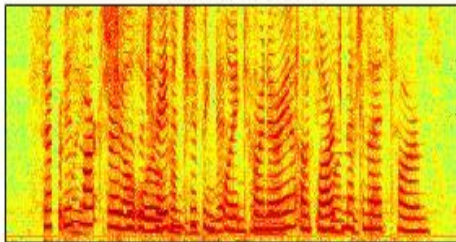
[1] M. Kolbek, Dong Yu, Z.-H. Tan and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol.25, No.10, pp.1901-1913, 2017

[2] C. Xu, W. Rao, E.-S. Chng, H. Li, "A Shifted Delta Coefficient Objective for Monaural Speech Separation Using Multi-task Learning," INTERSPEECH 2018: 3479-3483

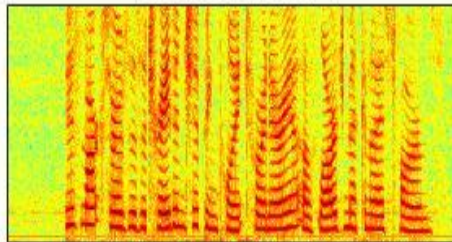
PIT Demo 1

Example: female-female speakers' mixture ('050a050i_2.1935_421c020b_-2.1935')

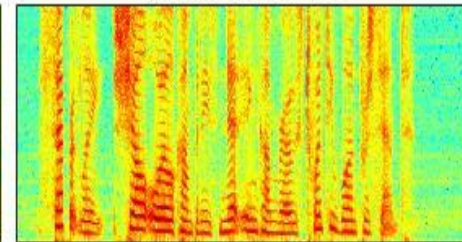
(a) Mixed Speech



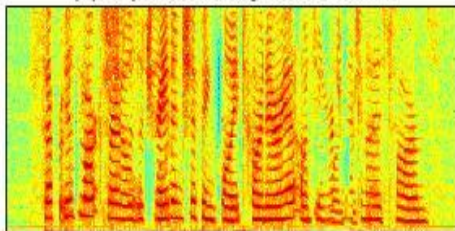
(b) Target Speaker 1



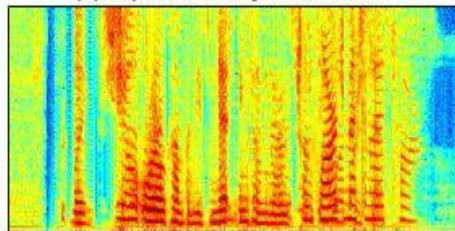
(c) Target Speaker 2



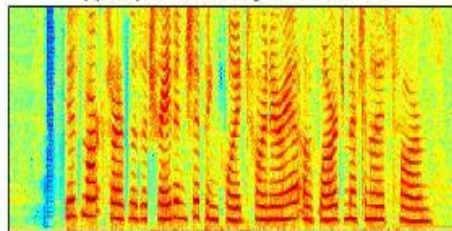
(d) Separation 1 by uPIT-BLSTM



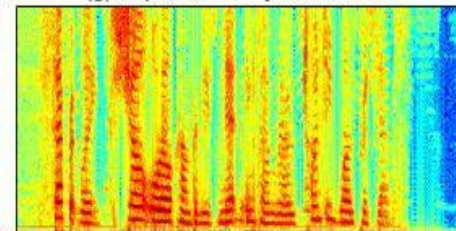
(e) Separation 2 by uPIT-BLSTM



(f) Separation 1 by SDC-G-MTL



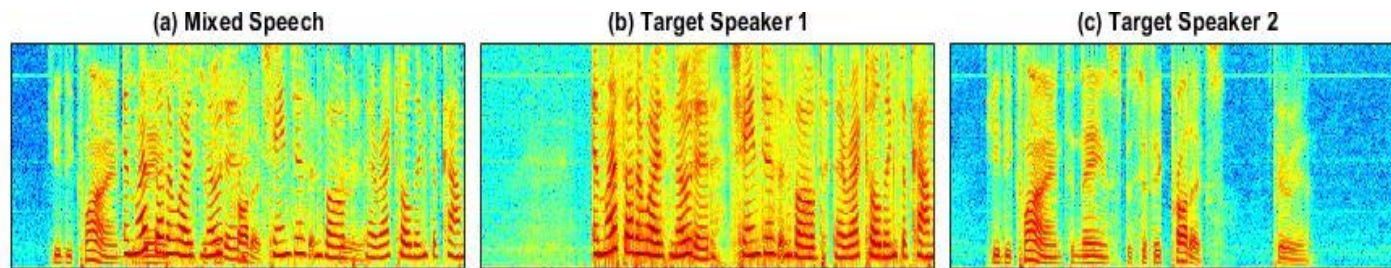
(g) Separation 2 by SDC-G-MTL



[1] Chenglin Xu, Wei Rao, Eng Siong Chng, Haizhou Li, "A Shifted Delta Coefficient Objective for Monaural Speech Separation Using Multi-task Learning," INTERSPEECH 2018: 3479-3483

PIT Demo 2

Example: male-female speakers' mixture ('441c020m_2.4506_447o030z_-2.4506')

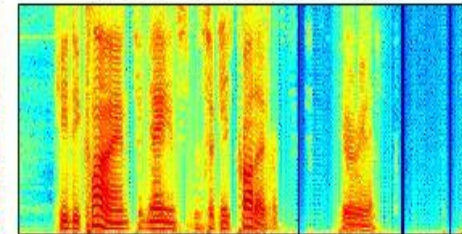
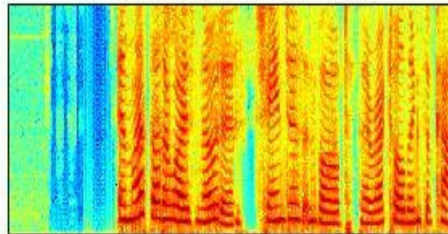
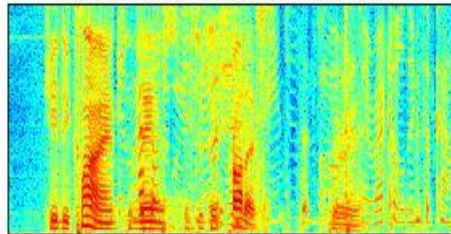
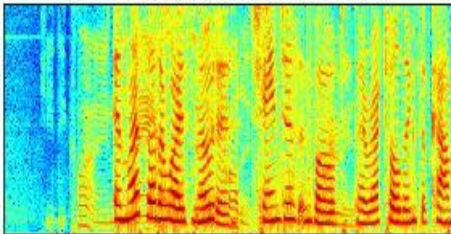


(d) Separation 1 by uPIT-BLSTM

(e) Separation 2 by uPIT-BLSTM

(f) Separation 1 by SDC-G-MTL

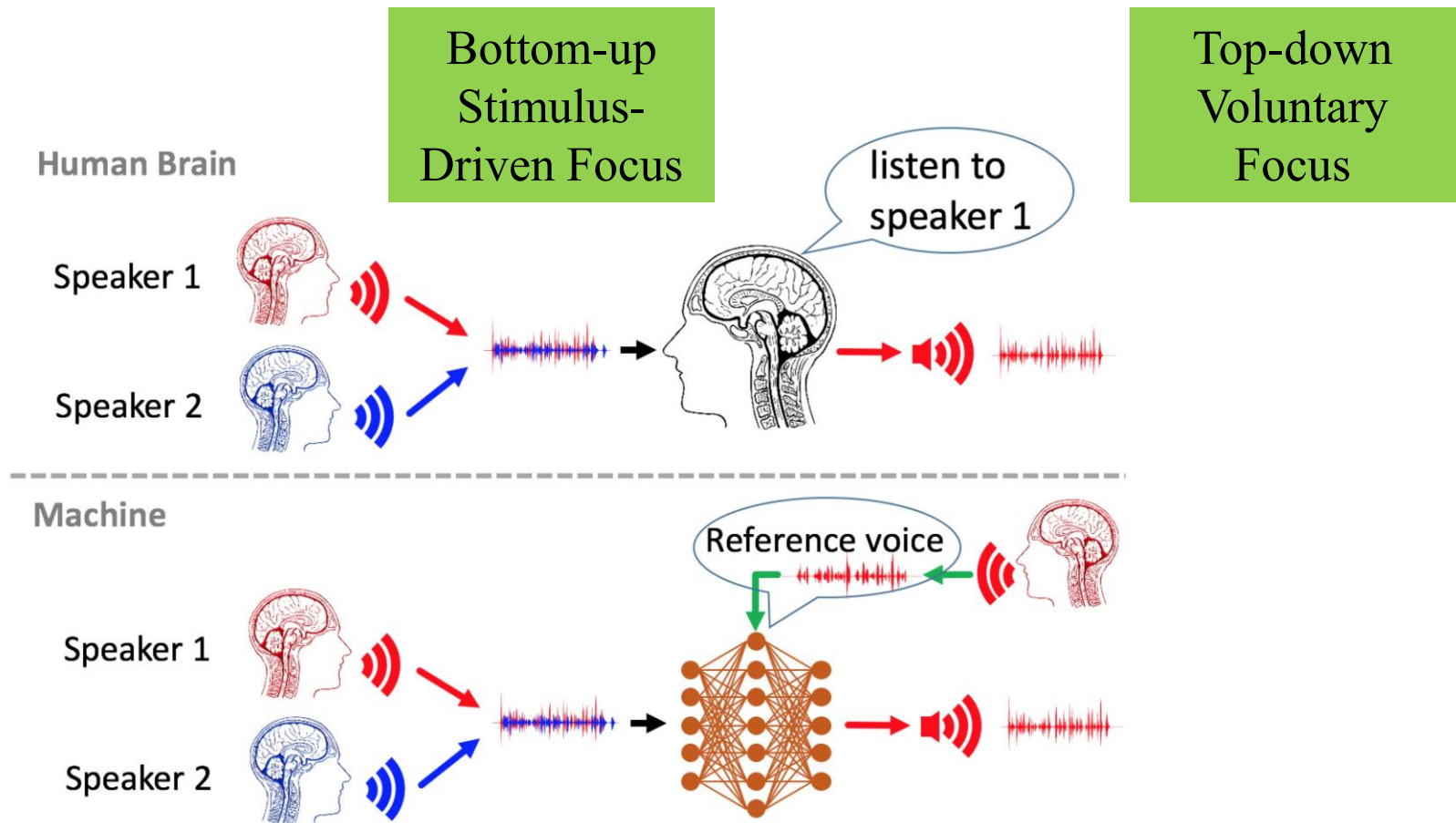
(g) Separation 2 by SDC-G-MTL



[1] Chenglin Xu, Wei Rao, Eng Siong Chng, Haizhou Li, "A Shifted Delta Coefficient Objective for Monaural Speech Separation Using Multi-task Learning," INTERSPEECH 2018: 3479-3483

Idea 4: SpEx I

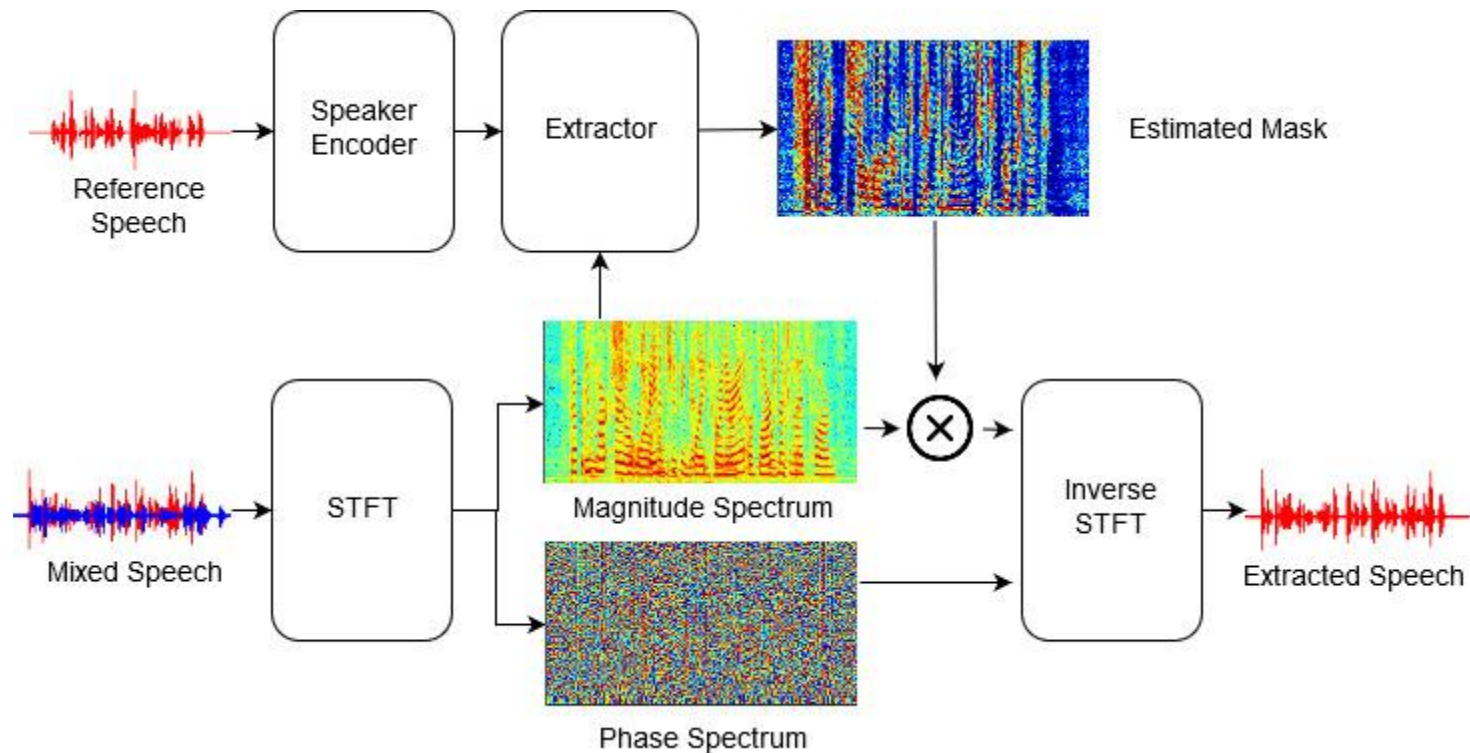
Frequency Domain Speaker Extraction



[1] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in ICASSP 2019 .

Idea 4: SpEx I

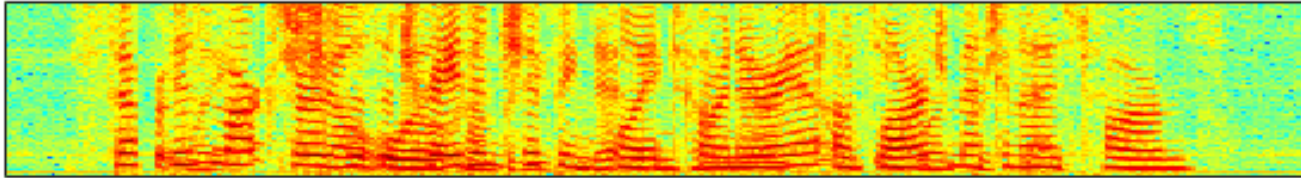
Frequency Domain Speaker Extraction



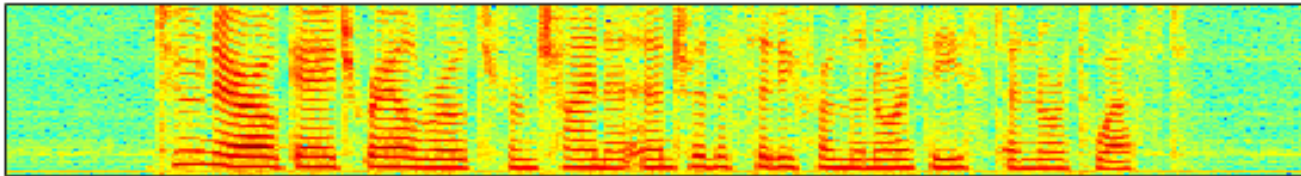
SpEx I - Demo 1

- Example: female-female speakers' mixture

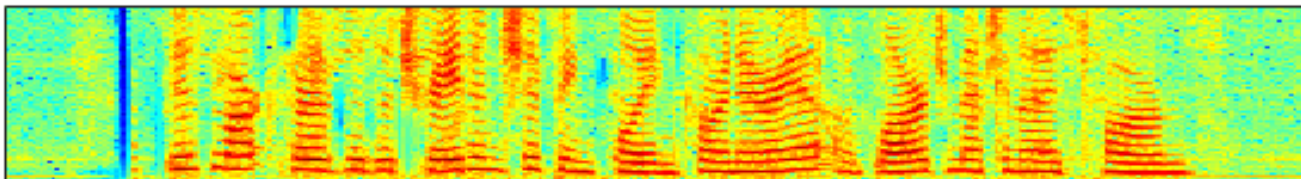
(a) Mixture



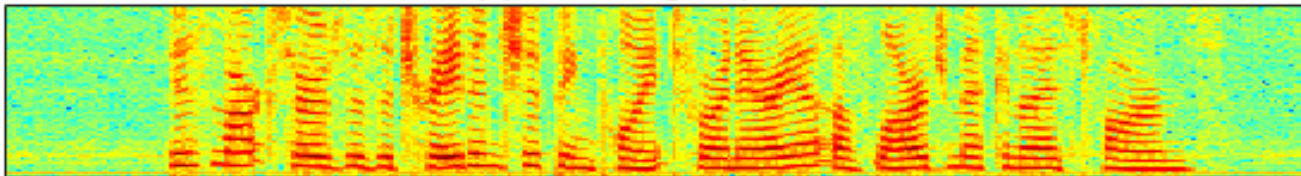
(b) Auxiliary Target Speaker



(c) Extracted Target Speaker



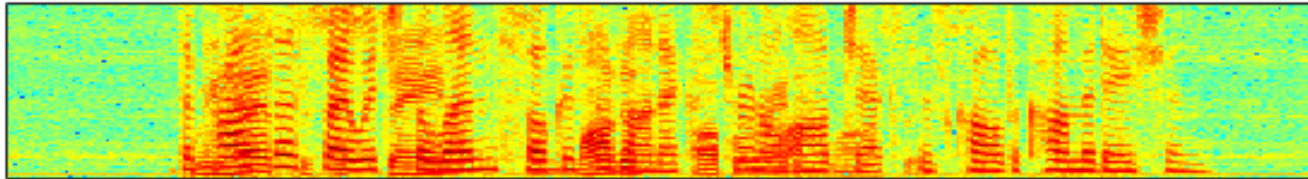
(d) Clean Target Speaker



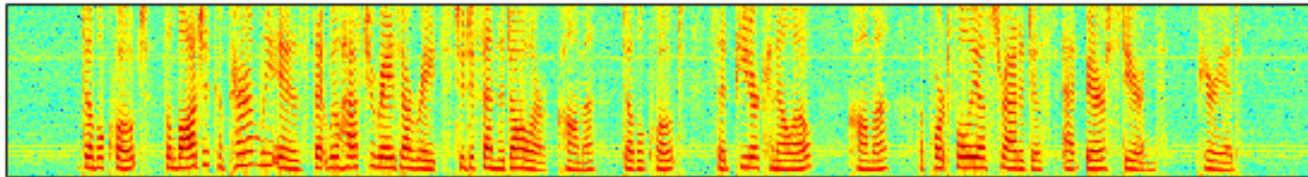
SpEx I - Demo 2

- Example: male-female speakers' mixture

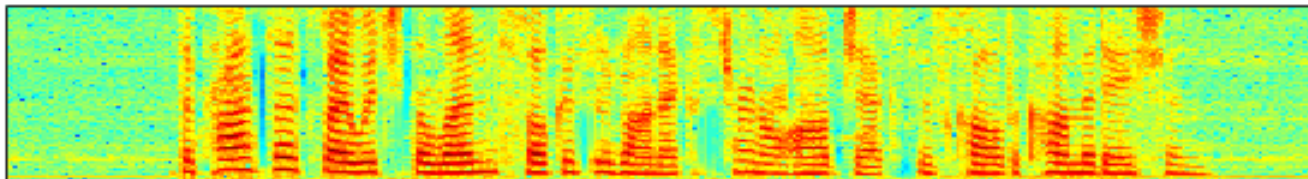
(a) Mixture



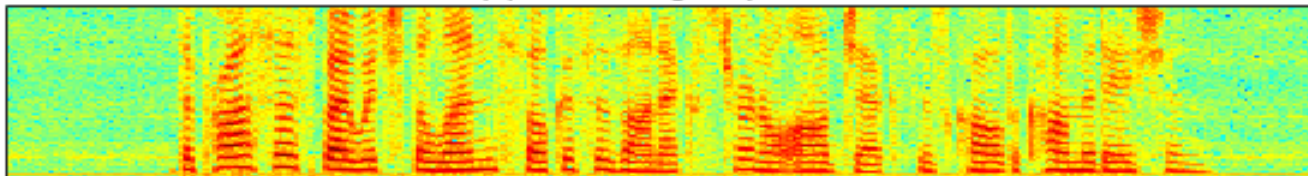
(b) Auxiliary Target Speaker



(c) Extracted Target Speaker

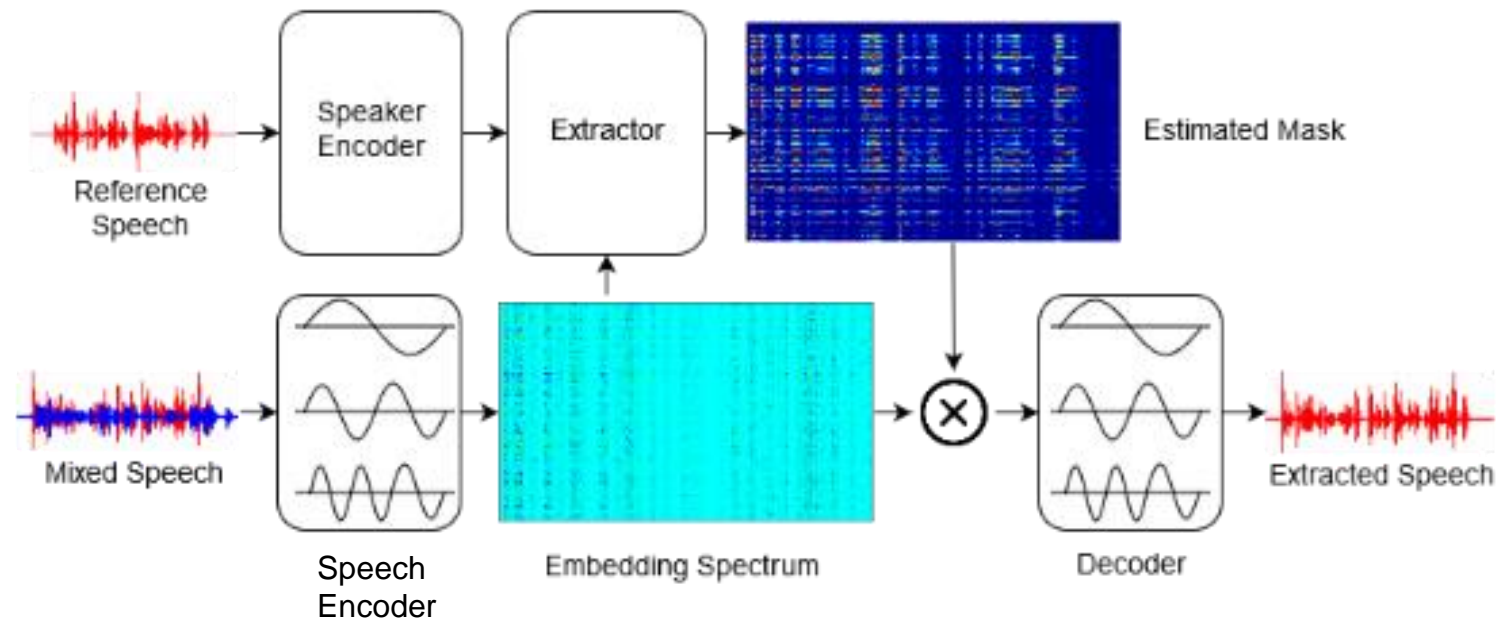


(d) Clean Target Speaker



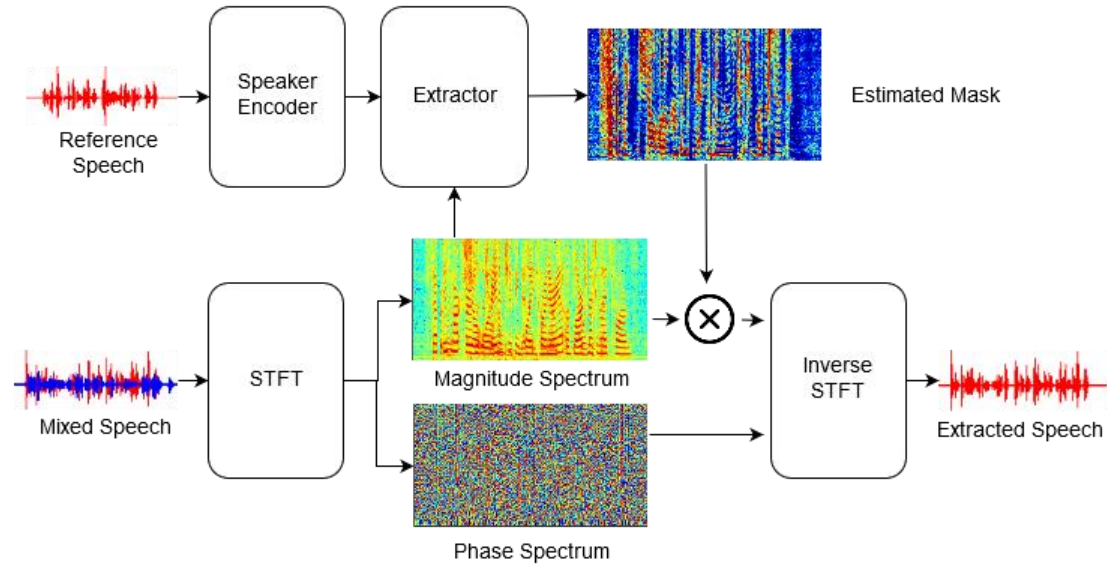
Idea 5: SpEx II

Time Domain Speaker Extraction

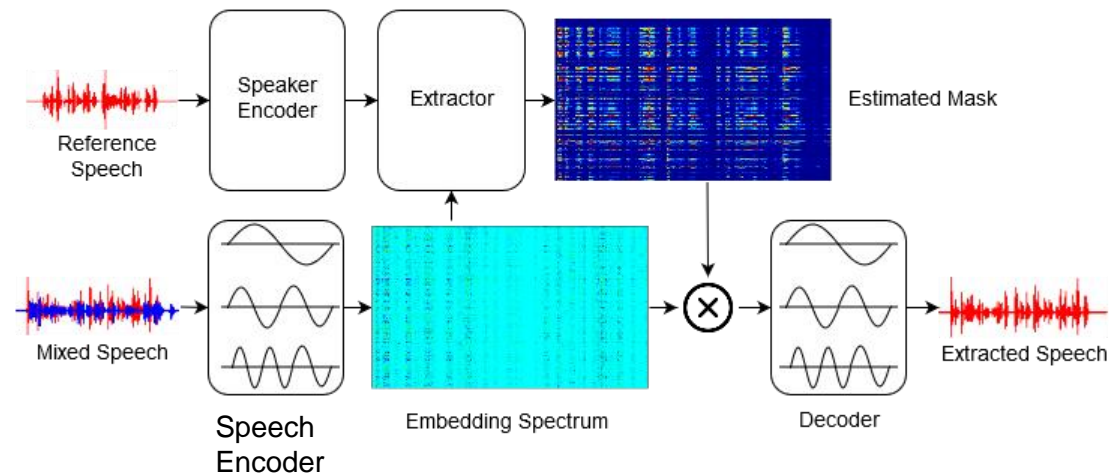


Frequency vs Time Domain

Frequency Domain Speaker Extraction



Time Domain Speaker Extraction











Idea 5: SpEx II Results

□ The results on WSJ0-2mix (max) database.

| Method | Paras | SDR | SI-SDR | PESQ |
|----------------------|-------|--------------|--------------|-------------|
| Mixture | - | 2.60 | - | 2.31 |
| SBF-IBM [1] | 19.3M | 6.45 | - | 2.32 |
| SBF-MSAL [1] | 19.3M | 9.62 | - | 2.64 |
| SBF-MTSAL [2] | 19.3M | 9.90 | - | 2.66 |
| SBF-MTSAL-Concat [2] | 8.9M | 10.99 | - | 2.73 |
| SpEx II | 9.0M | 12.78 | 12.19 | 2.92 |

- [1] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. Of ICASSP*. IEEE, 2018, pp. 5554-5558.
- [2] C. Xu, W. Rao, E. S. Chng, and H. Li, "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss", in *Proc. Of ICASSP*. IEEE, 2019, pp. 6990-6994.

Idea 5: SpEx II Demo (Time domain)

| Type | Mixture | Auxiliary | Clean | Extracted |
|---------------|---|--|---|---|
| Male-Female |  |  |  |  |
| Female-Female |  |  |  |  |

Agenda

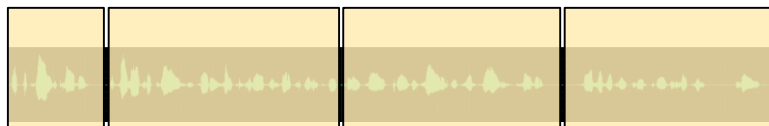
- Selective auditory attention
- Speech separation & speaker extraction
- **Applications**

Temporal vs Spectral Speaker Diarization

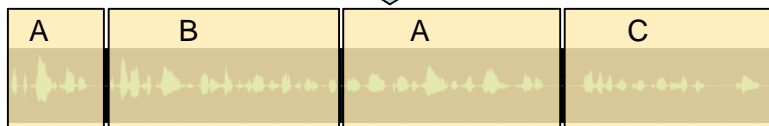
Multi-talker Speech:



Segmentation:

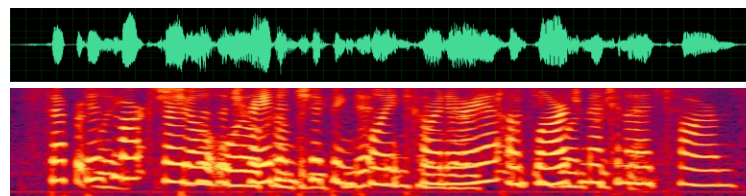


Clustering by Speaker:

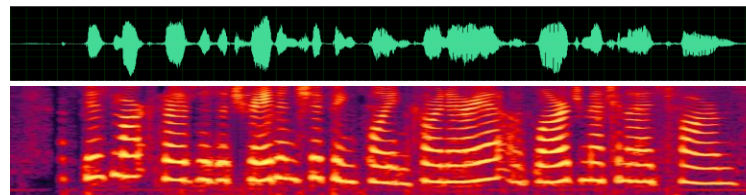


Speaker diarization [1,2] –
temporal segregation

Multi-talker Speech:



Speaker A:



Speaker extraction –
spectral segregation

[1] David Snyder et. al, "Speaker recognition for multi-speaker conversations using x-vectors", ICASSP 2019.

[2] Rohan Kumar Das et. al, "Speaker clustering with penalty distance for speaker verification with multi-speaker speech", APSIPA ASC 2019.

Speech Separation + Speech Recognition

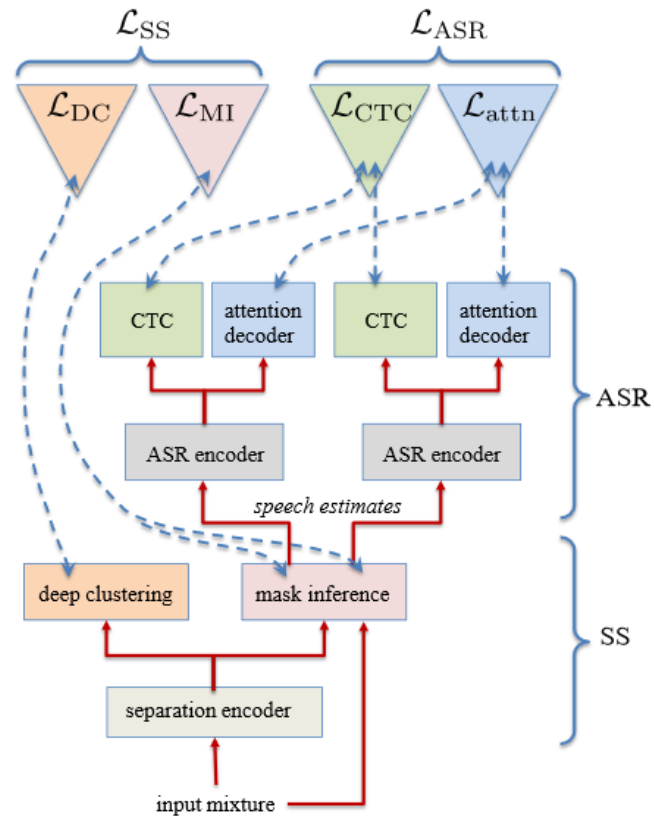
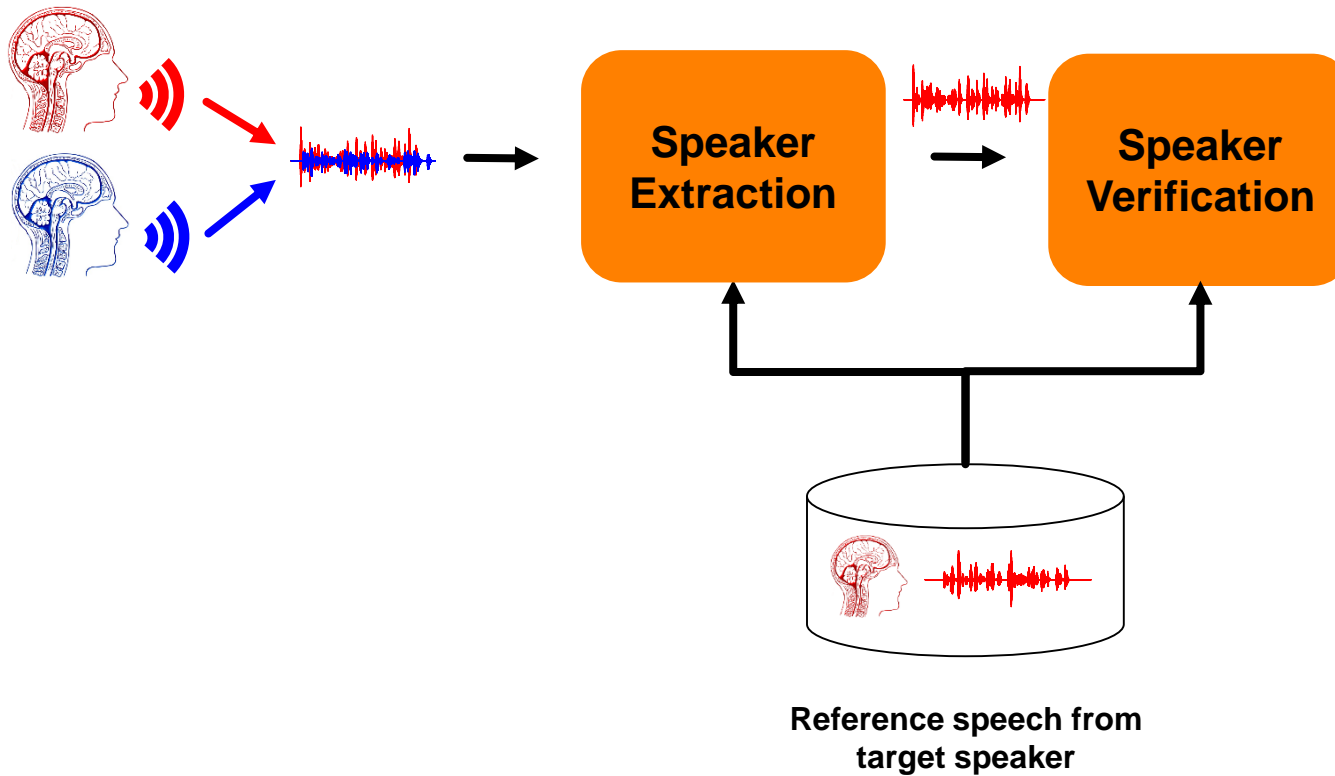
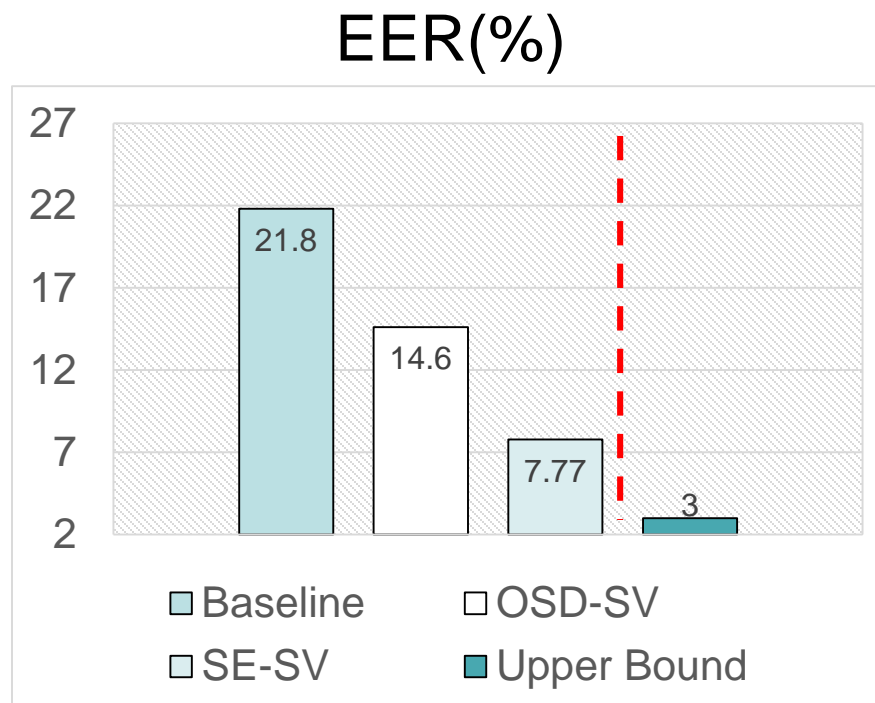


Fig. 1. End-to-end separation and recognition network

Speaker Extraction + Speaker Verification



Speaker Extraction + Speaker Verification



Upper bound: clean test speech for evaluation, only one speaker in the speech.

- SE-SV significantly improves the performance of multi-talker SV and achieve 64.4% relative EER reduction over the zero-effort baseline.
- SE-SV significantly outperforms oracle speaker diarization (OSD) in the overlapped multi-talker scenarios.

Thank you!



Human brain is more power efficient than our most efficient computer by orders of magnitude. It is vital to draw inspiration from how human brain works.