Recent Advances in Speaker Extraction

Haizhou Li

National University of Singapore, Singapore

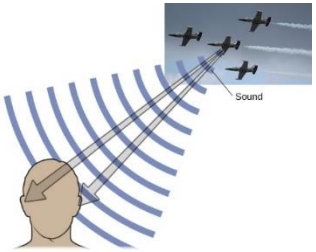
Agenda

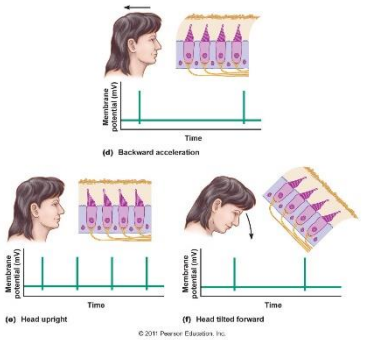
- **Selective auditory attention**
- **Speaker extraction**
- **ICASSP Paper Review**

Ears and Hearing

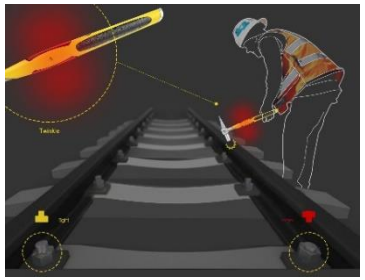
Frequency Analyser

Localization/Equilibrium





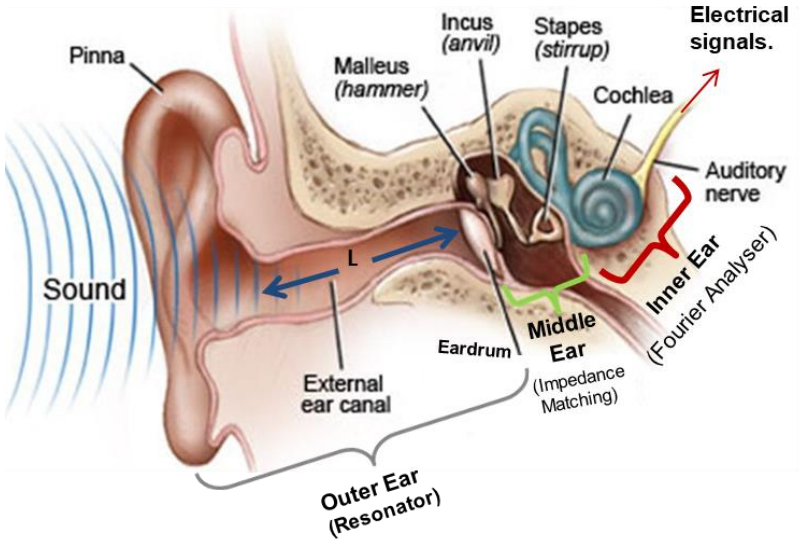
Attention



From the Perspective of Physiology

(生理学)

Human Auditory System

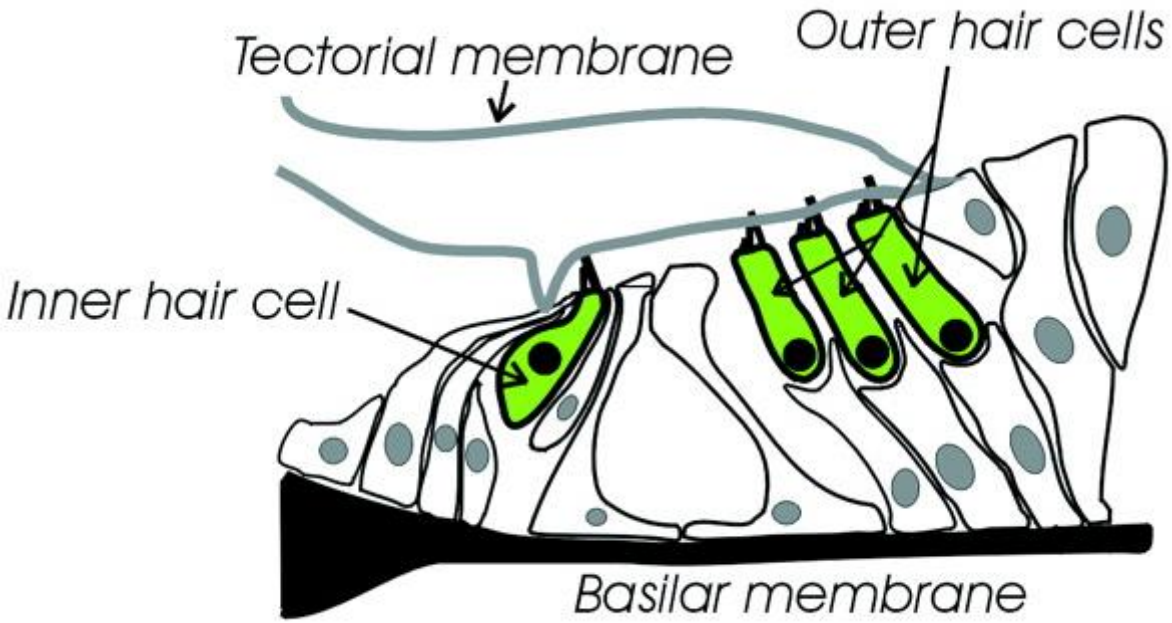


✓ The human outer ear is most sensitive at about 3kHz and provides about 20dB (decibels) of gain to the eardrum at around 3kHz.

✓ 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.

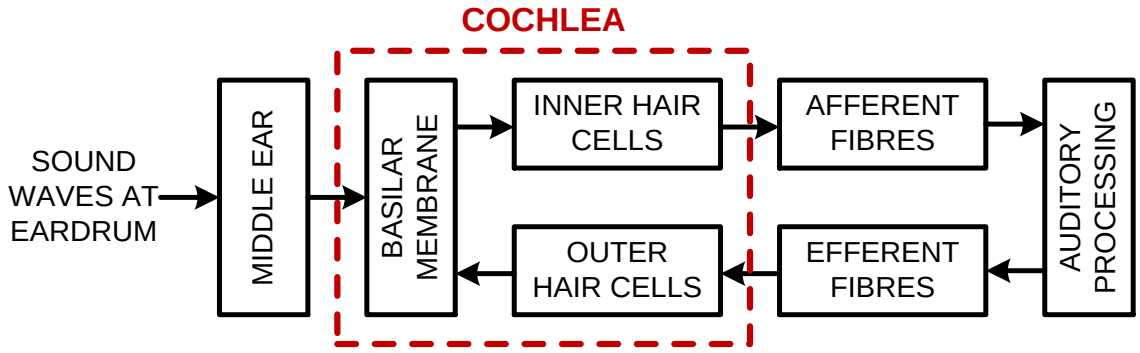
Structure of the Cochlea



- 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

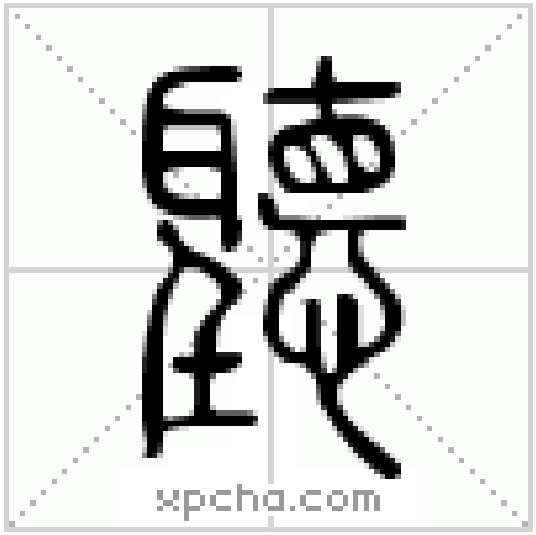
From the Perspective of Neuroscience

(神经科学)

Chinese Character ‘to listen’

Feature Extraction and Acoustic Modeling (the ear)

A person

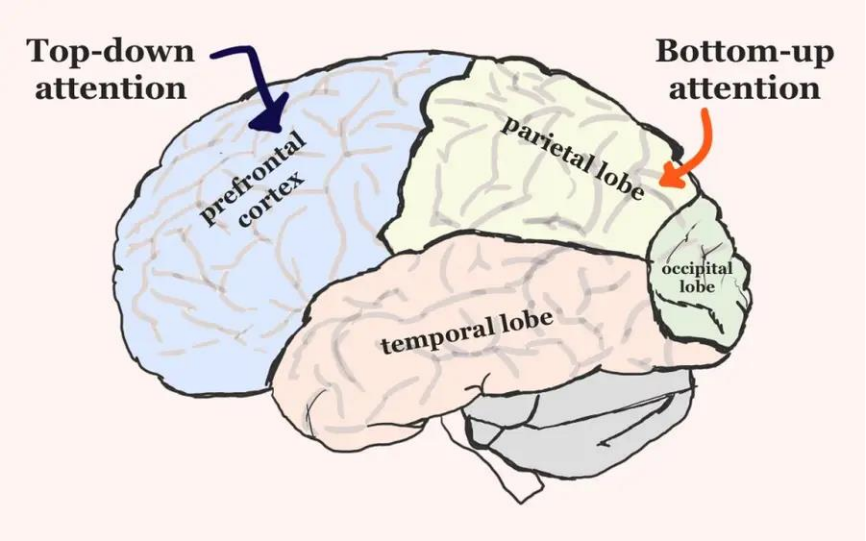


Undivided attention (the eyes)

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**

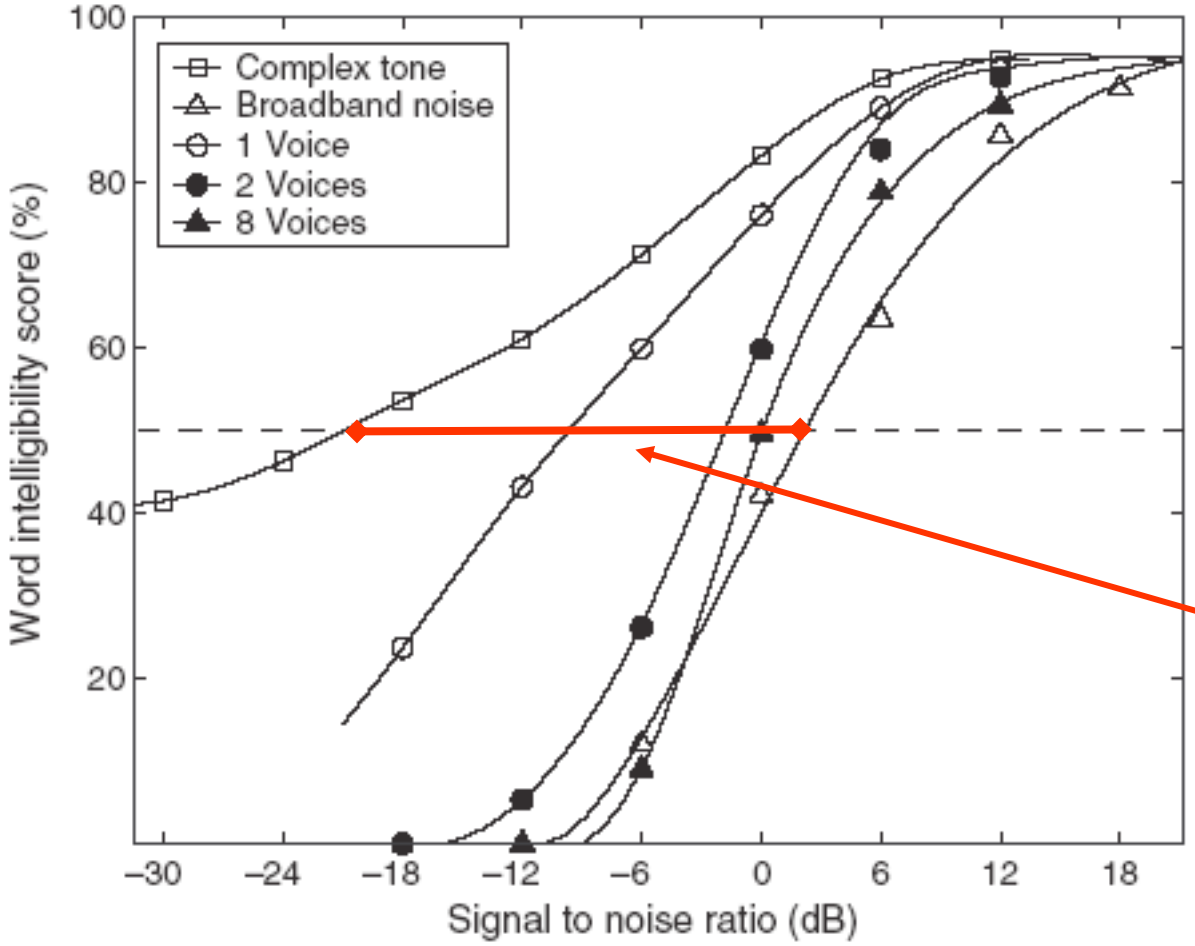


Timothy J. Buschman, Earl K. Miller, Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices, Science 315(5820):1860-2 · March 2007

From the Perspective of Psychoacoustics

(心理声学)

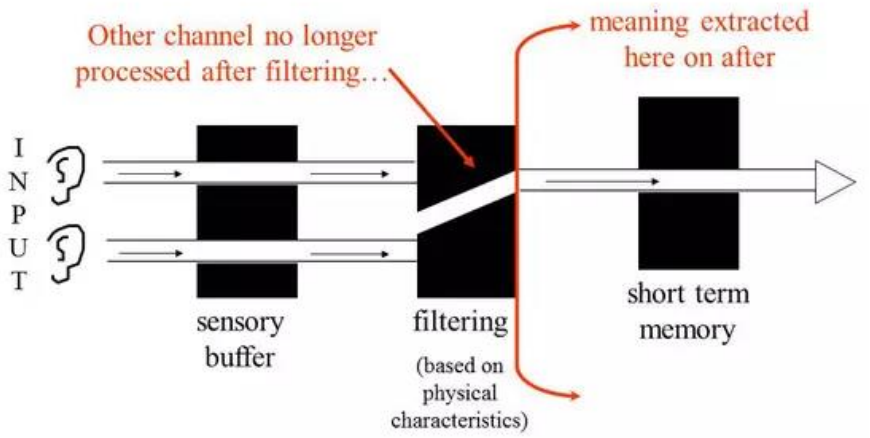
Effects of Competing Source



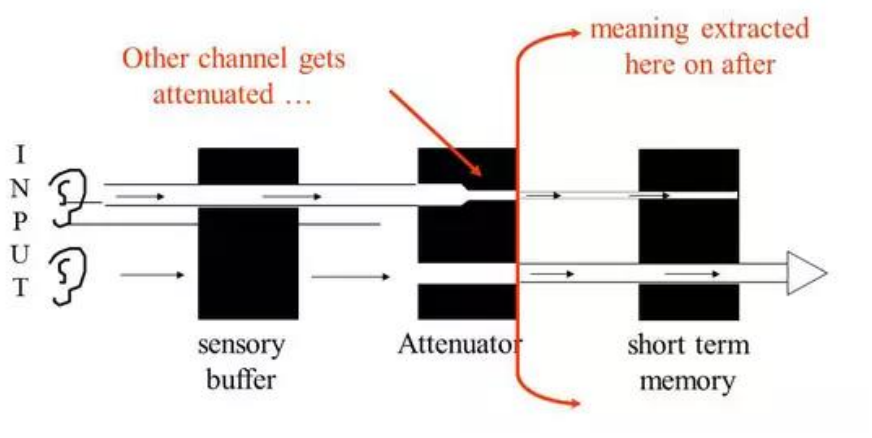
Speech Reception Threshold Difference (23 dB!)

G. A. Miller, "The masking of speech," *Psychol. Bull.*, vol. 44, pp. 105–129, 1947.

Theory of Filtering



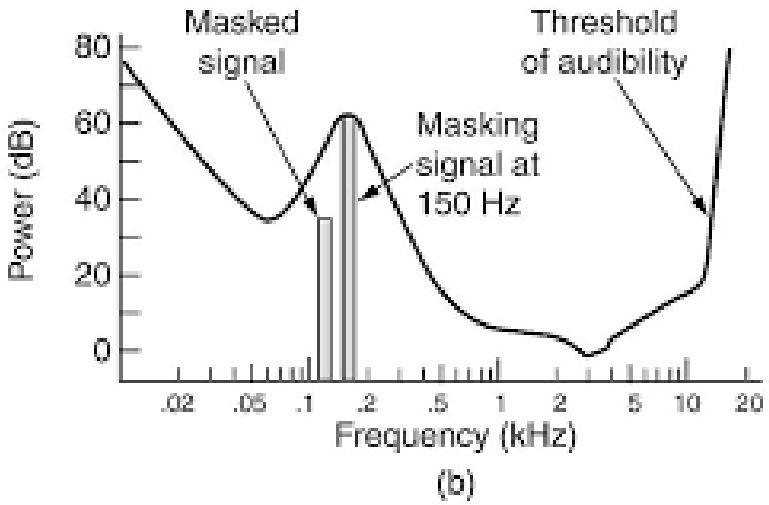
Broadbent (1958)



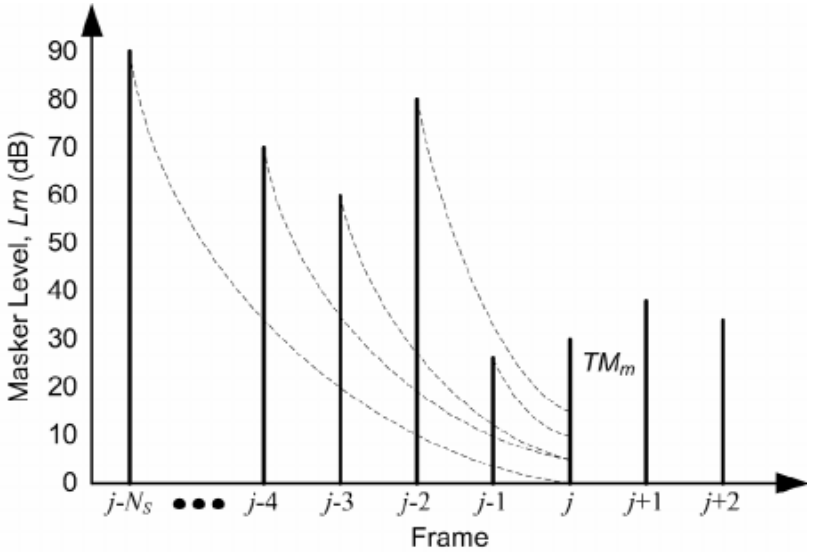
Treisman (1964)

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

NUS National University of Singapore

From the Perspective of Signal Processing

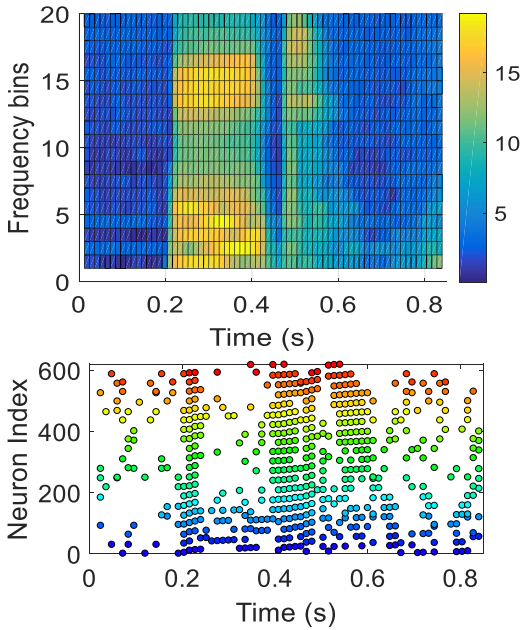
(信号处理)

Auditory Masking

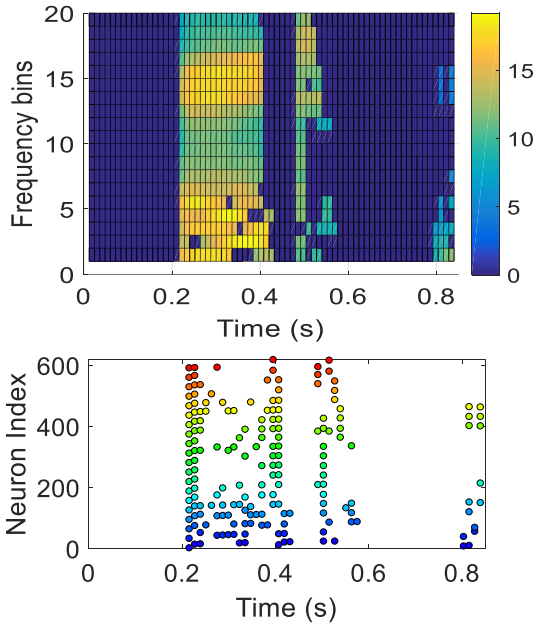
- **Example (not the best): Ideal Binary Mask (IBM)**
  - $s(t,f)$: Target energy in unit (t,f)
  - $n(t,f)$: Noise energy
  - θ: A local SNR criterion (LC) in dB, typically chosen to be 0 dB
  - It does not actually separate the mixture!

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

Masking Effect



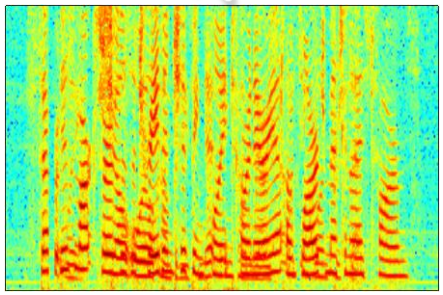
Original



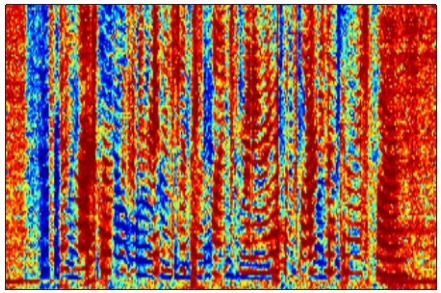
After Masking

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

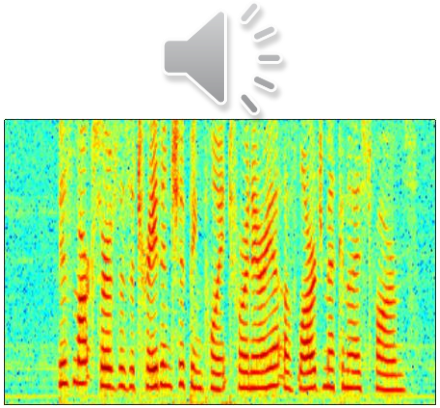
Masking Effect (Ideal Ratio Mask)



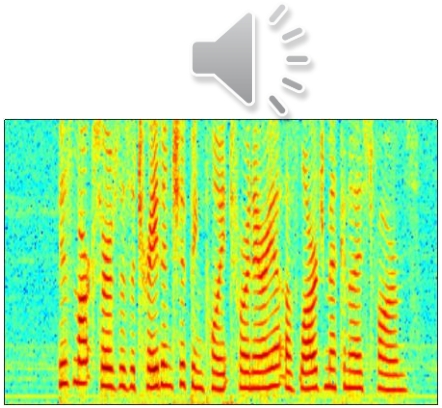
Female-Female Mixed Speech



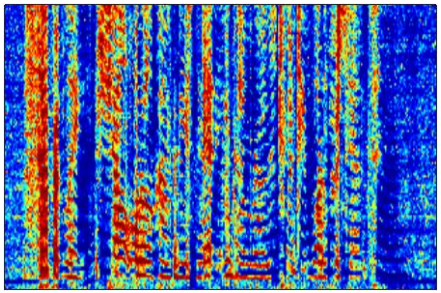
IRM for Speaker 1



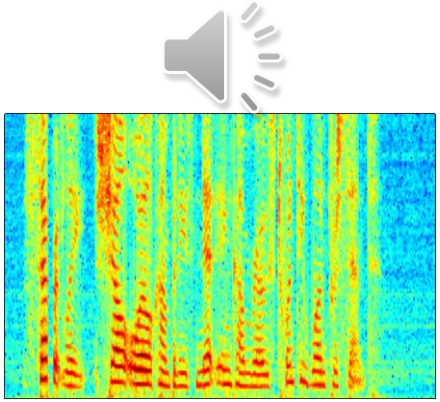
Separated Speech of Speaker 1 with IRM



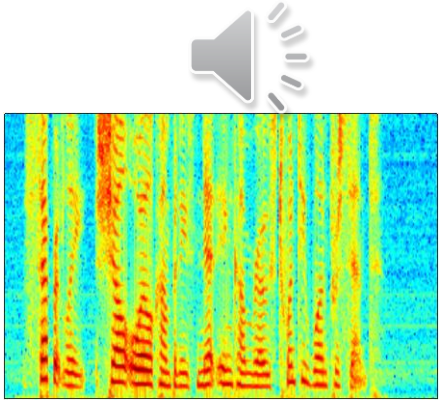
Clean Speech of Speaker 1



IRM for Speaker 2

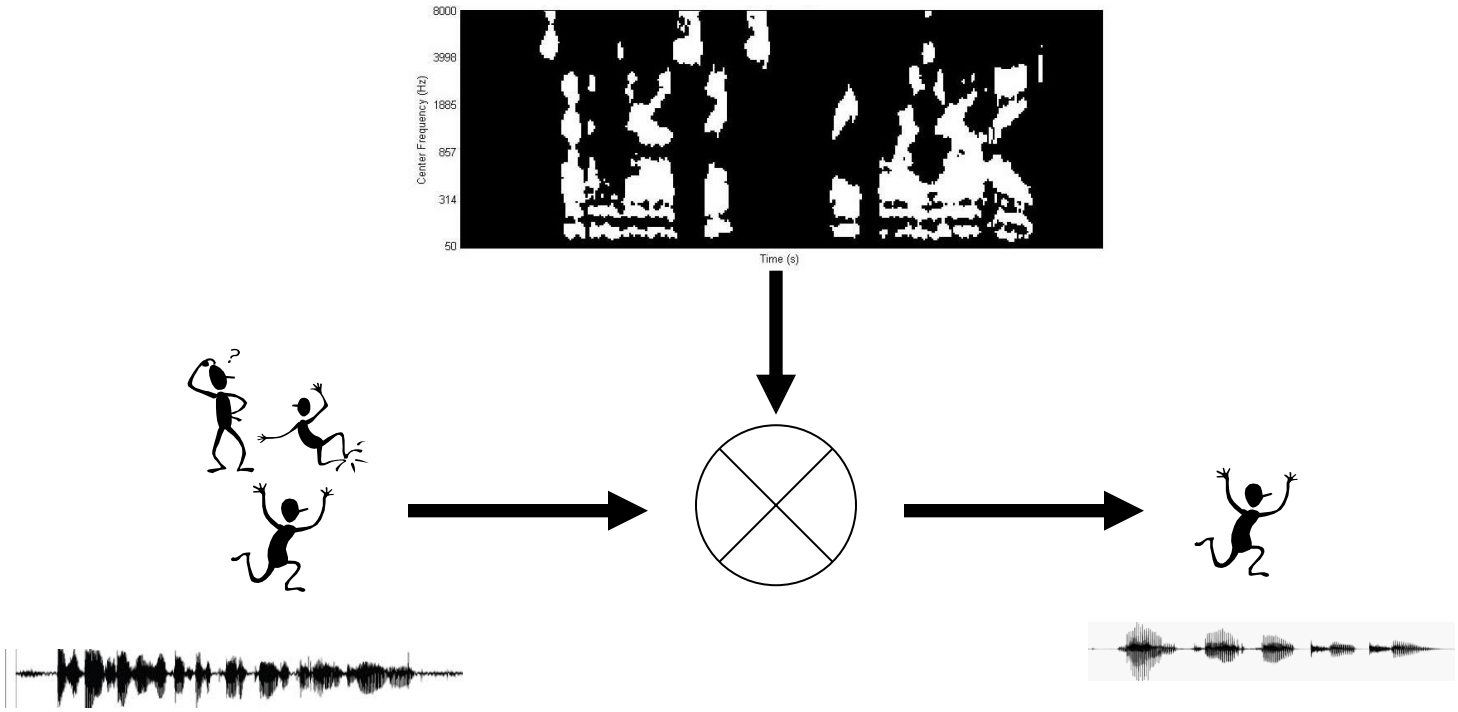


Separated Speech of Speaker 2 with IRM



Clean Speech of Speaker 2

Question: How to find the mask?

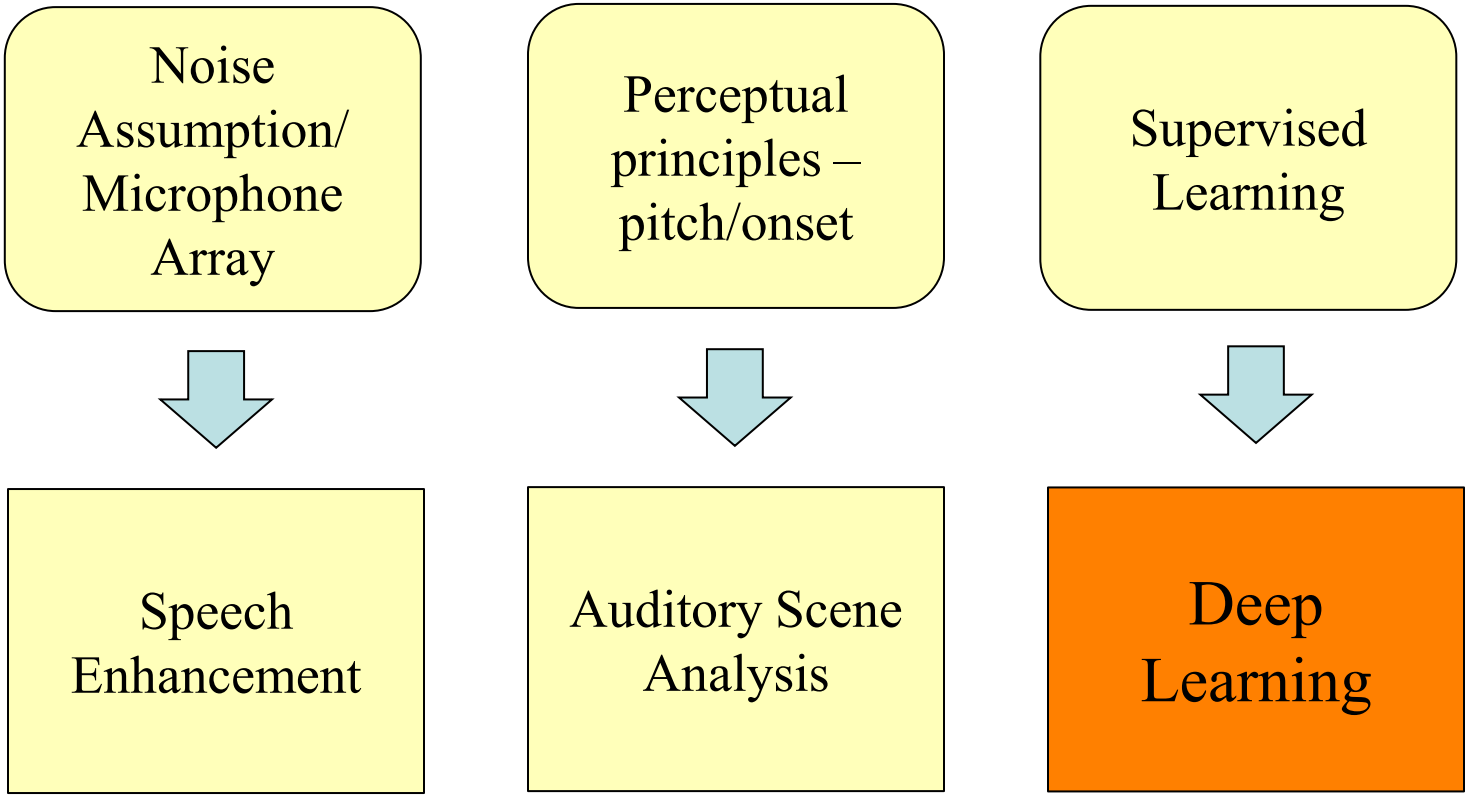


Spectro-Temporal Receptive Fields that reflect the temporal and spectral modulations belonging to the target sound events

Agenda

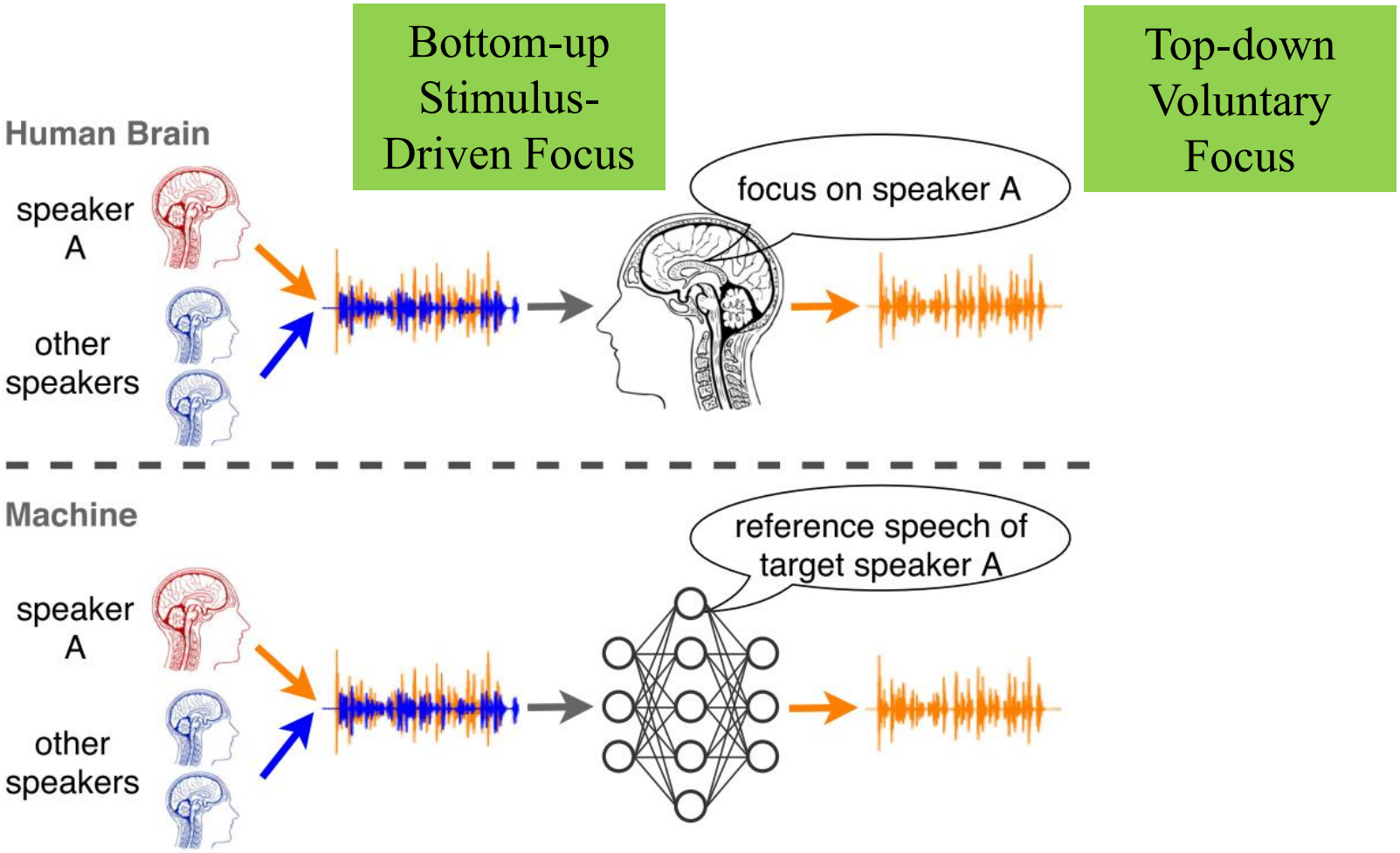
- Selective auditory attention
- **Speaker extraction**
- ICASSP Paper Review

Monaural Speech Separation: Overview



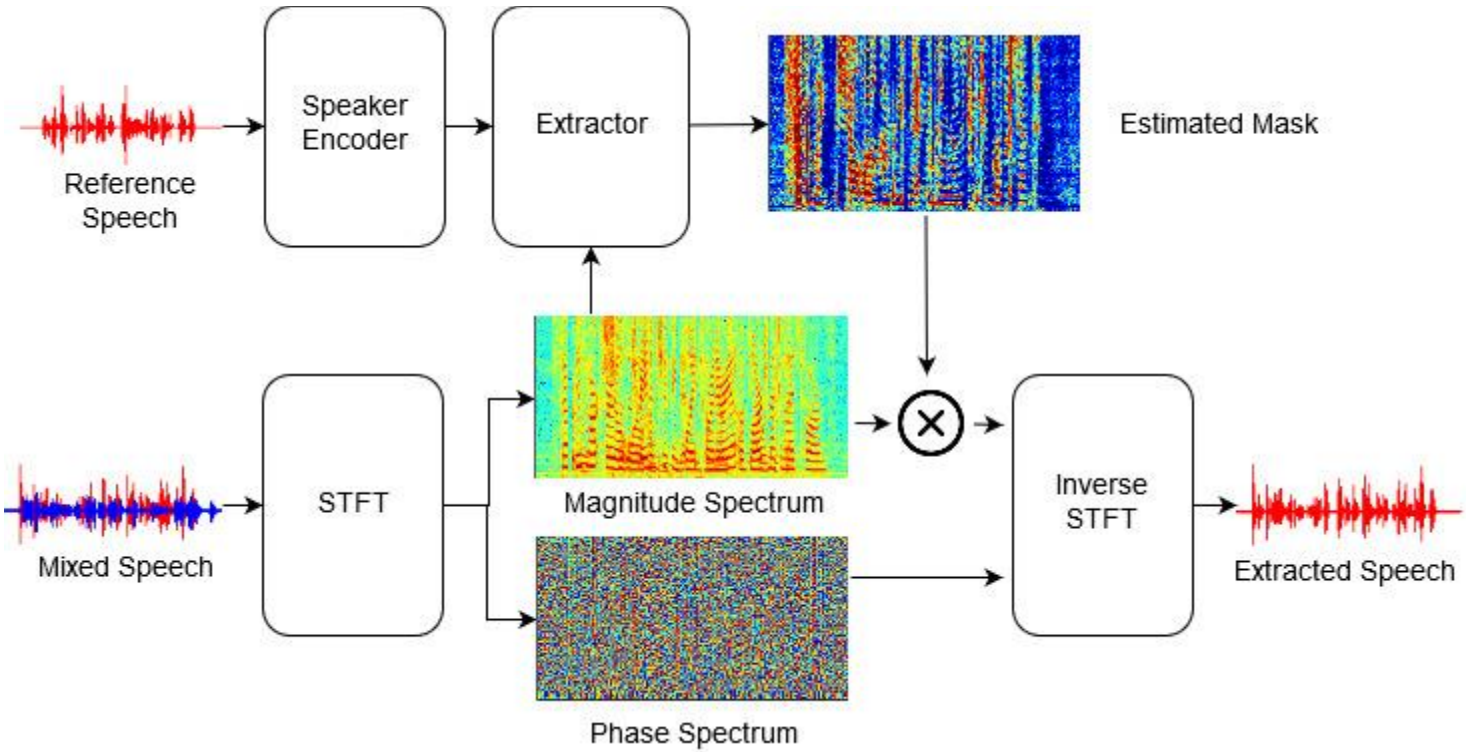
[1] Albert S. Bregman, Auditory Scene Analysis. Cambridge, MA, USA: MIT Press, 1990.
[2] D. Wang, Supervised Speech Separation Based on Deep Learning: An Overview, IEEE/ACM T-ASLP 2018

Selective auditory attention by Speaker Extraction



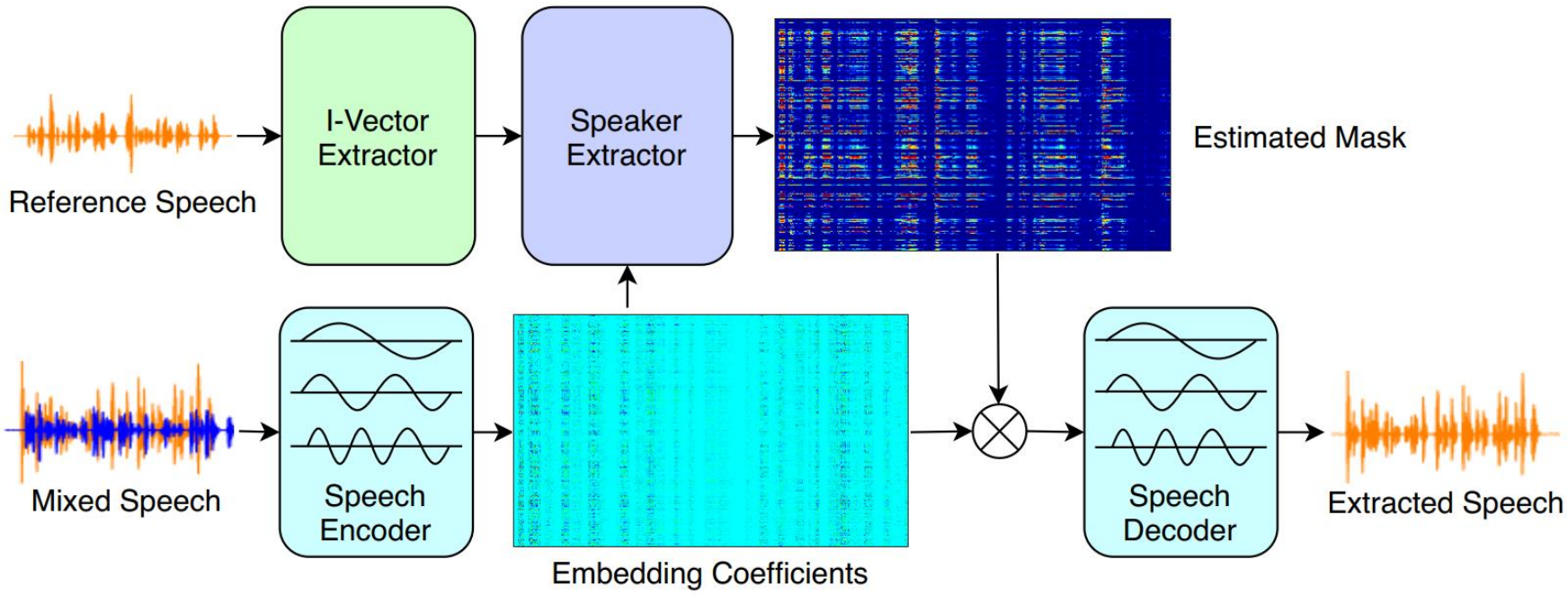
SBF-MTSAL-Concat

Frequency Domain Speaker Extraction



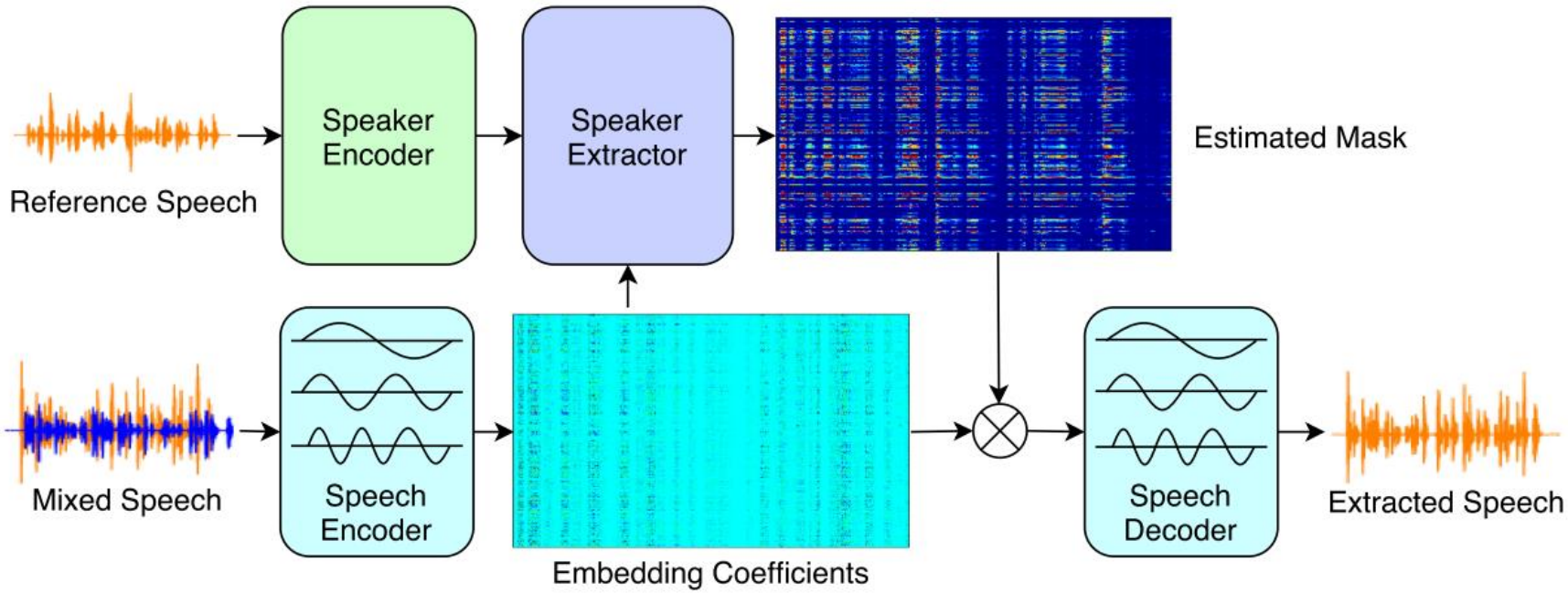
[1] Chenglin Xu, Wei Rao, Eng Siong Chng, and Haizhou Li, “Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss,” in ICASSP 2019 .

TseNet
Time Domain Speaker Extraction



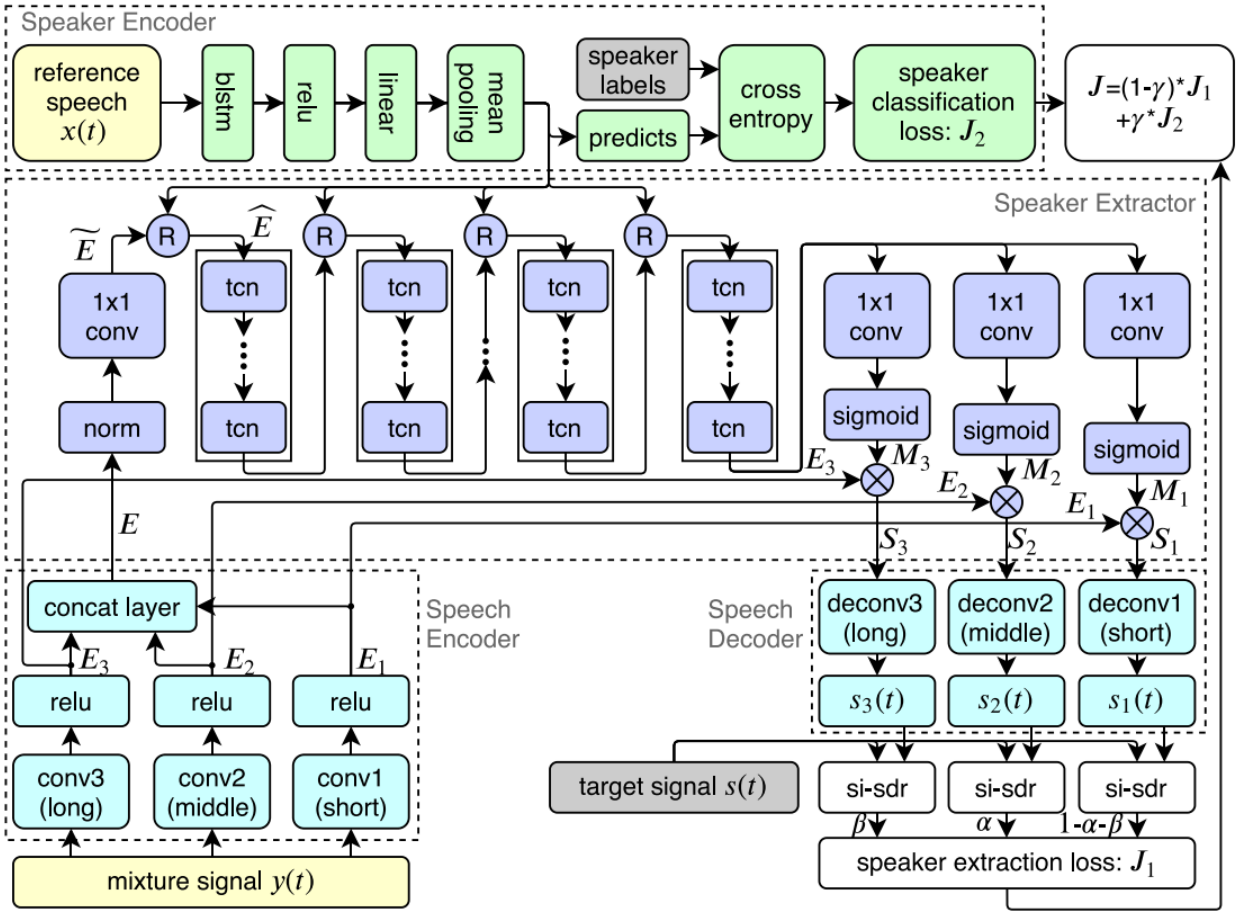
Chenglin Xu, Wei Rao, Eng Siong Chng, Haizhou Li, “Time-domain speaker extraction network”, ASRU 2019.

SpEx
Time Domain Speaker Extraction



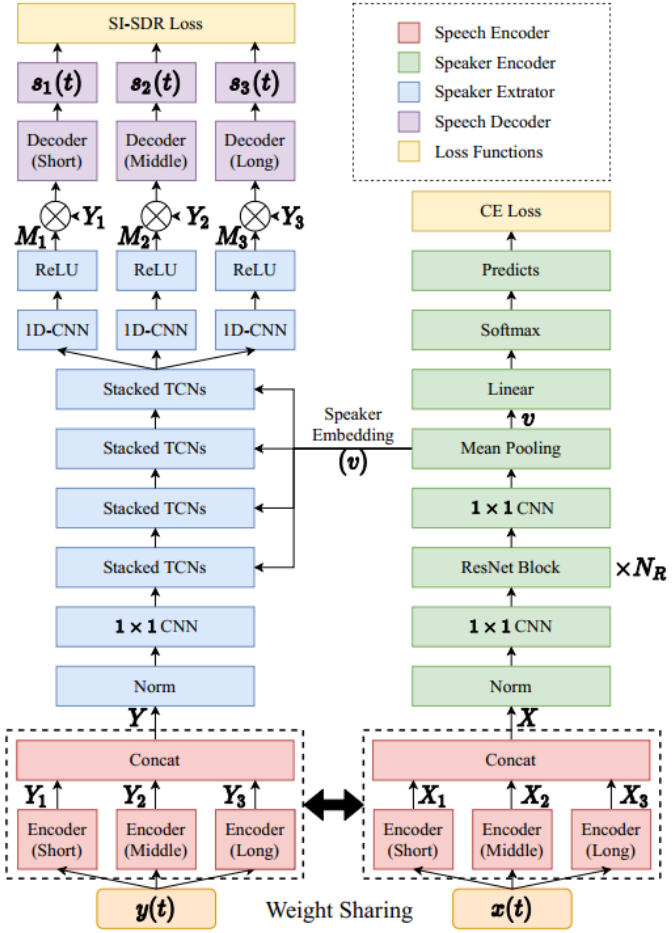
Chenglin Xu, Wei Rao, Eng Siong Chng, Haizhou Li, “SpEx: multi-scale time-domain speaker extraction network”, IEEE/ACM Transaction on Audio, Speech and Language Processing, 2020.

SpEx
Time Domain Speaker Extraction



Chenglin Xu, Wei Rao, Eng Siong Chng, Haizhou Li, "SpEx: multi-scale time-domain speaker extraction network", IEEE/ACM Transaction on Audio, Speech and Language Processing, 2020.

SpEx+
Time Domain Speaker Extraction



Meng Ge, Chenglin Xu, Longbiao Wang, Eng Siong Chng, Jianwu Dang, and Haizhou Li. "SpEx+: A Complete Time Domain Speaker Extraction Network." arXiv preprint arXiv:2005.04686 (2020).

Results

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

Methods	Domain	SDR	SI-SDR	PESQ
Mixture	-	2.60	2.50	2.31
SpeakerBeam [1]	Freq-	9.62	9.22	2.64
SBF-MTSAL-Concat [2]	Freq-	11.39	10.60	2.77
TseNet [3]	Time	15.24	14.73	3.14
SpEx	Time	17.15	16.68	3.36
SpEx+	Time	**18.54**	**18.20**	**3.49**

[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.
[3] C. Xu, W. Rao, E. S. Chng, H. Li, "Time-domain speaker extraction network", ASRU 2019..

Results

- **The results on WSJ0-2mix (min) database.**

Task	Methods	#Params	SDRi	SI-SDR
BSS	DPCL++ [4]	13.6M	-	10.8
	uPIT-BLSTM-ST [5]	92.7M	10.0	-
	DANet [6]	9.1M	-	10.5
	cuPIT-Grid-RD [7]	53.2M	10.2	-
	SDC-G-MTL [8]	53.9M	10.5	-
	CBLDNN-GAT [9]	39.5M	11.0	-
	Chimera++ [10]	32.9M	12.0	11.5
	WA-MISI-5 [11]	32.9M	13.1	12.6
BSS	BLSTM-TasNet [12]	23.6M	13.6	13.2
BSS	Conv-TasNet [13]	5.1M	15.6	15.3
BSS	DPRNN-TasNet [14]	2.6M	19.0	18.8
SE	SpEx	10.8M	17.0	16.6
SE	SpEx+	13.3M	17.6	17.4

References

[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.
[3] C. Xu, W. Rao, E. S. Chng, H. Li, "Time-domain speaker extraction network", ASRU 2019..
[4] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," arXiv preprint arXiv:1607.02173, 2016
[5] M. Kolbæk, D. Yu, Z. 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.
[6] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 246–250.
[7] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 6–10.
[8] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "A shifted delta coefficient objective for monaural speech separation using multitask learning," in Interspeech, 2018, pp. 3479–3483.
[9] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, "CBLDNN-based speaker-independent speech separation via generative adversarial training," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 711–715
[10] Z. Wang, J. L. Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 686–690
[11] Z. Wang, J. L. Roux, D. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," arXiv preprint arXiv:1804.10204, 2018.
[12] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network." in Interspeech, 2018, pp. 342–346.
[13] 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.
[14] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 46–50.

Agenda

- Selective auditory attention
- Speaker extraction
- **ICASSP Paper Review (speech separation)**

ICASSP 2020 Paper Review I

NUS National University of Singapore

A MULTI-PHASE GAMMATONE FILTERBANK FOR SPEECH SEPARATION VIA TASNET

David Ditter and Timo Gerkmann

Signal Processing (SP), Universität Hamburg, Germany
david.ditter@uni-hamburg.de, timo.gerkmann@uni-hamburg.de

ABSTRACT

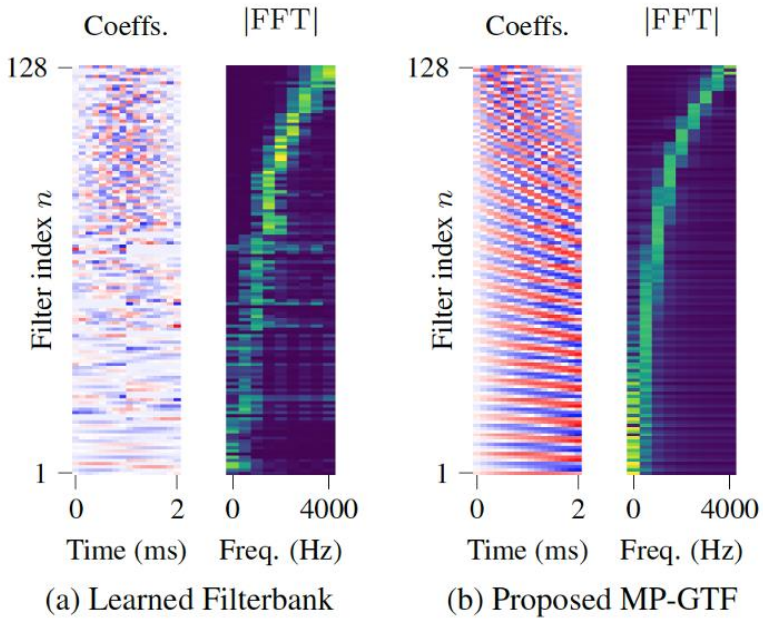
In this work, we investigate if the learned encoder of the end-to-end convolutional time domain audio separation network (Conv-TasNet) is the key to its recent success, or if the encoder can just as well be replaced by a deterministic hand-crafted filterbank. Motivated by the resemblance of the trained encoder of Conv-TasNet to auditory filterbanks, we propose to employ a deterministic gammatone filterbank. In contrast to a common gammatone filterbank, our filters are restricted to 2 ms length to allow for low-latency processing. Inspired by the encoder learned by Conv-TasNet, in addition to the logarithmically spaced filters, the proposed filterbank holds multiple gammatone filters at the same center frequency with varying phase shifts. We show that replacing the learned encoder with our proposed multi-phase gammatone filterbank (MP-GTF) even leads to a scale-invariant source-to-noise ratio (SI-SNR) improvement of 0.7 dB. Furthermore, in contrast to using the learned encoder we show that the number of filters can be reduced from 512 to 128 without loss of performance.

finally, at least for the convolutional time domain audio separation network (Conv-TasNet) and FurcaNext, the separation section of the network is implemented as a temporal convolutional network with a bottleneck structure [10] instead of an architecture using Long Short-Term Memory (LSTM) layers. In [7], this modification has empirically shown to give better average results and to improve the robustness of these systems against time shifts of the input signal.

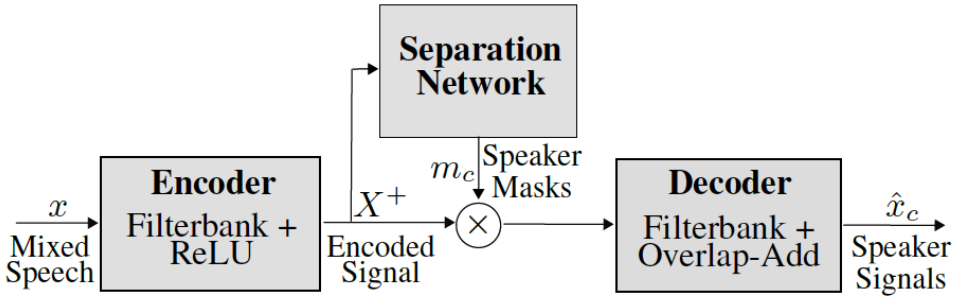
When replacing the deterministic STFT analysis-synthesis structure by a learned encoder-decoder structure, the following general question arises: Should we use a well-understood, deterministic encoder (analysis filterbank) which is based on signal processing principles and possibly motivated by perceptual features? Or should we let the network run free and find a data-driven signal encoding for the given problem all by itself? This question has very recently gained attention and was investigated in several research papers such as [11], [12] and [13].

On a theoretical level, there are good arguments for both choices. Advocating for a learned encoder, we can argue that we might obtain a better network after training if the network has a high degree

ICASSP 2020 Paper Review I



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



ICASSP 2020 Paper Review I

Encoder	Decoder	N	SI-SNRi (dB)	
			Train	Test
Learned	Learned	512	19.3	15.4
MP-GTF	Learned	512	19.0	**15.9**
MP-GTF	MP-GTF Pseudo Inv.	512	18.1	15.4

Table 2. SI-SNR improvements on WSJ0-MIX2 training and test set for different configurations of the encoder and decoder of Conv-TasNet for $N = 512$ filters. Higher is better.

Encoder	Decoder	N	SI-SNRi (dB)
Learned	Learned	**512**	**15.4**
Learned	Learned	128	15.2
MP-GTF	Learned	512	15.9
MP-GTF	Learned	**128**	**16.1**
MP-GTF	Learned	64	15.4
MP-GTF	Learned	48	14.4

Table 3. SI-SNR improvements on WSJ0-MIX2 test set for different values of the number of filters N with a learned decoder.

ICASSP 2020 Paper Review II

NUS National University of Singapore

FILTERBANK DESIGN FOR END-TO-END SPEECH SEPARATION

Manuel Pariente [1], *Samuele Cornell* [2], *Antoine Deleforge* [1], *Emmanuel Vincent* [1]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2]Department of Information Engineering, Università Politecnica delle Marche, Italy

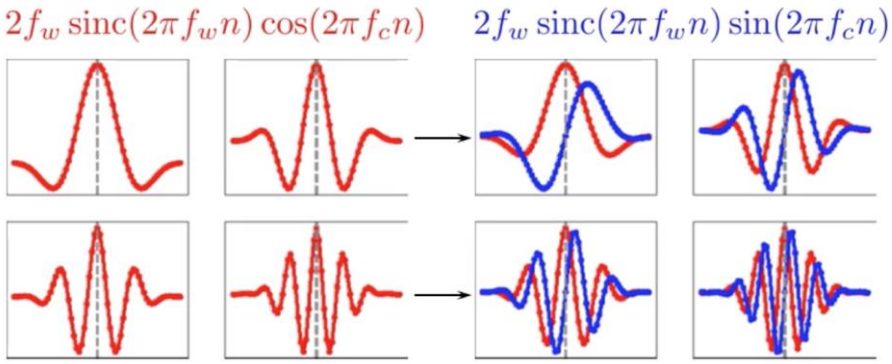
ABSTRACT

Single-channel speech separation has recently made great progress thanks to learned filterbanks as used in ConvTasNet. In parallel, parameterized filterbanks have been proposed for speaker recognition where only center frequencies and bandwidths are learned. In this work, we extend real-valued learned and parameterized filterbanks into complex-valued analytic filterbanks and define a set of corresponding representations and masking strategies. We evaluate these filterbanks on a newly released noisy speech separation dataset (WHAM). The results show that the proposed analytic learned filterbank consistently outperforms the real-valued filterbank of ConvTasNet. Also, we validate the use of parameterized filterbanks and show that complex-valued representations and masks are beneficial in all conditions. Finally, we show that the STFT achieves its best performance for 2 ms windows.

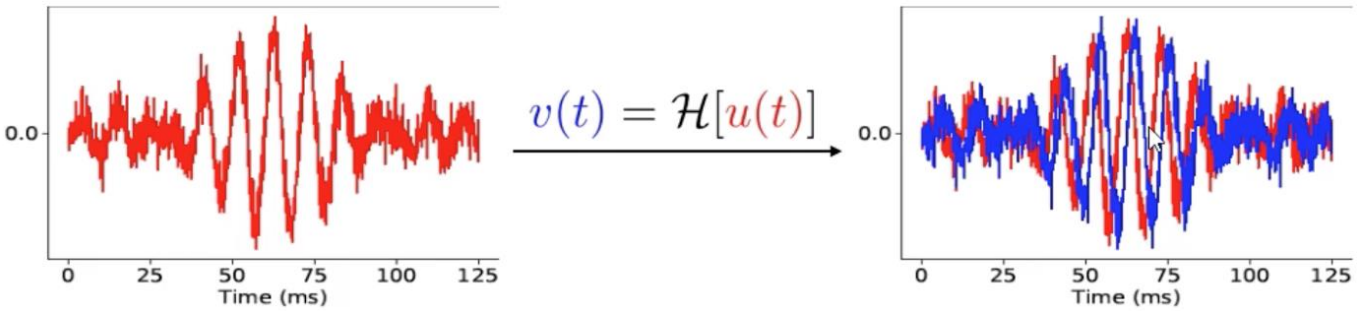
jointly with the masking network [7–10]. While learned representations have been shown to be undeniably superior to the STFT for speech separation in clean recording conditions [7–9], their impact in the presence of noise has been lesser studied. In fact, the authors of [14] introduce a noisy extension of wsj0-2mix, WHAM, on which initial results indicate that the advantage of learned representations reduces as noise is introduced, suggesting that learning from the raw waveform might be harder in noisy conditions. In parallel, parameterized kernel-based filterbanks have been introduced as a front-end for speech and speaker recognition [15, 16]. The underlying idea is to restrict the filters to a certain family of functions and jointly learn their parameters with the network. These filterbanks are meant for signal analysis only, though.

In this paper, we define suitable parameterized filters for analysis-synthesis. Compared with fixed STFT filters, the proposed filters offer more flexibility and diversity thanks to

ICASSP 2020 Paper Review II



$u_{\text{analytic}}(t) = u(t) + j\mathcal{H}[u(t)]$, $\mathcal{H}$ is the Hilbert transform (quadrature phase shift)



ICASSP 2020 Paper Review II

Window size (ms)	2	5	10	25	50
Param.	2.3	1.0	0.6	-0.8	-2.7
Param.(3x filters)	2.3	1.2	0.7	-0.7	-2.7
Param.+$\mathcal{H}$	**11.8**	**11.6**	**9.1**	**7.3**	**4.0**

Table 1. SI-SDR$_i$ (dB) as a function of window size for parametric filterbanks in clean conditions. Bold values represent the best statistically significant results.

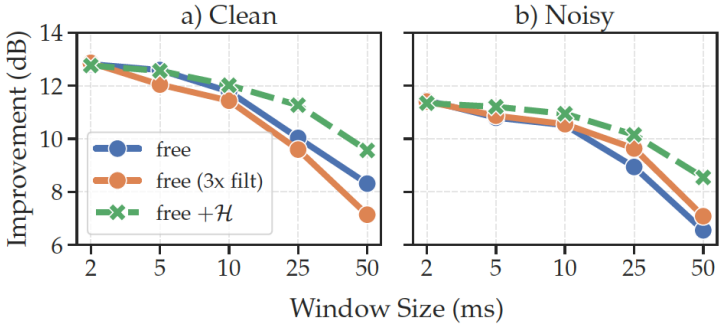


Fig. 1. SI-SDR$_i$ as a function of window size for free filterbanks in clean and noisy conditions.

Model	Dataset	separate-clean	separate-noisy
chimera++ [6]	8kHz min	11.0	9.9
Conv-TasNet [9][3]	8kHz min	15.1	**12.7**
Free+$\mathcal{H}$	8kHz min	**15.8**	**12.9**
chimera++ [6]	16kHz max	9.6	10.2
Conv-TasNet [9]	16kHz max	13.6	13.3
Free+$\mathcal{H}$	16kHz max	**14.0**	**14.0**

Table 3. SI-SDR$_i$ (dB) comparison between the proposed analytic free filterbank and previously proposed models. Bold values represent the best statistically significant results.

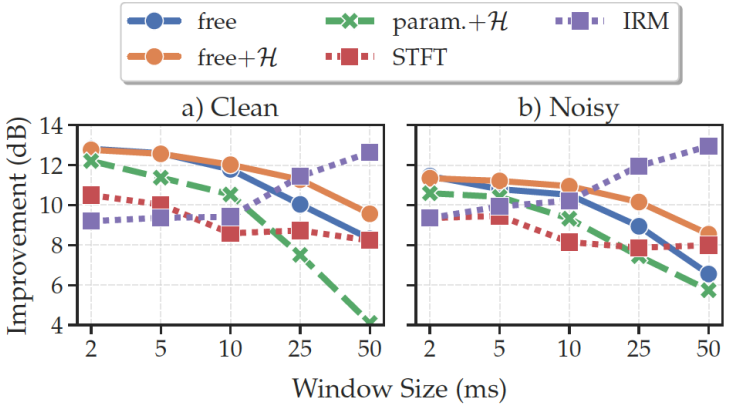


Fig. 3. SI-SDR$_i$ as a function of window size for all filterbanks and for the ideal ratio mask (IRM).

ICASSP 2020 Paper Review III

NUS National University of Singapore

TWO-STEP SOUND SOURCE SEPARATION: TRAINING ON LEARNED LATENT TARGETS

Efthymios Tzinis[♮] *Shrikant Venkataramani*[♮] *Zhepei Wang*[♮] *Cem Subakan*[♭] *Paris Smaragdis*[♮,♯]

[♮] University of Illinois at Urbana-Champaign
[♭] Mila–Quebec Artificial Intelligence Institute
[♯] Adobe Research

ABSTRACT

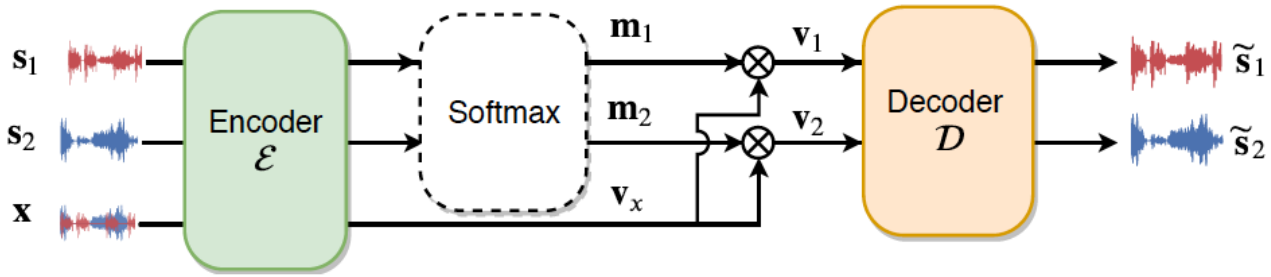
In this paper, we propose a two-step training procedure for source separation via a deep neural network. In the first step we learn a transform (and it's inverse) to a latent space where masking-based separation performance using oracles is optimal. For the second step, we train a separation module that operates on the previously learned space. In order to do so, we also make use of a scale-invariant signal to distortion ratio (SI-SDR) loss function that works in the latent space, and we prove that it lower-bounds the SI-SDR in the time domain. We run various sound separation experiments that show how this approach can obtain better performance as compared to systems that learn the transform and the separation module jointly. The proposed methodology is general enough to be applicable to a large class of neural network end-to-end separation systems.

not always yield an optimal decomposition of the input mixtures resulting to worse performance than the fixed STFT bases [17].

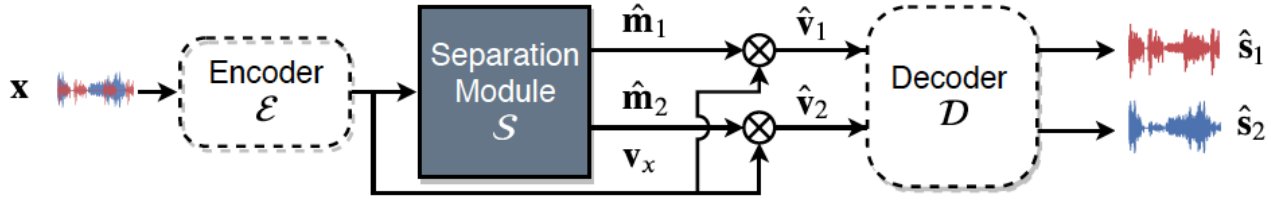
Some studies have reported significant benefits when performing source-separation in two stages. In [18], first the sources are separated and in a second stage the interference between the estimated sources is reduced. Similarly, an iterative scheme is proposed in [17], where the separation estimates from the first network are used as input to the final separation network. In [19], speaker separation is performed by first separating frame-level spectral components of speakers and later sequentially grouping them using a clustering network. Lately, state-of-the-art results in most natural language processing tasks have been achieved by pre-training the encoder transformation network [20].

In this work, we propose a general two-step approach for performing source separation which can be used in any mask-based

ICASSP 2020 Paper Review III



(a) Step 1: Learning the latent targets.

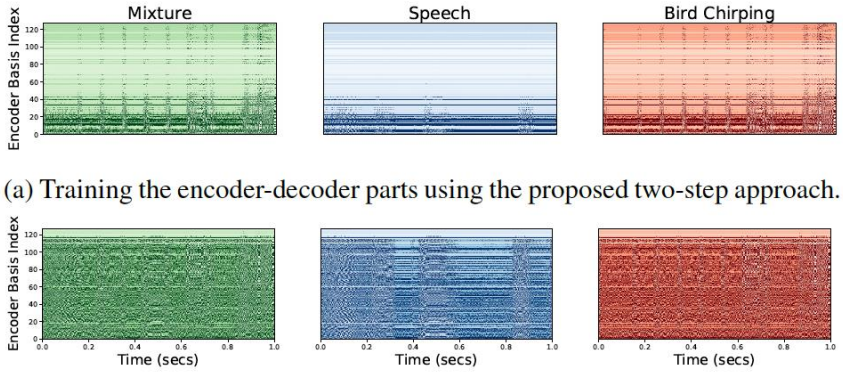


(b) Step 2: Training the separation module to produce the latent targets.

ICASSP 2020 Paper Review III

Separation Module	Target Domain	Sound Separation Task		
		Speech	Non-speech	Mixed
TDCN	Time	15.4	7.7	11.7
	Latent	16.1	8.2	12.4
RTDCN	Time	15.6	8.3	12.0
	Latent	16.2	8.4	12.6
Oracle Masks	STFT	13.0	14.8	14.5
	Latent	34.1	39.2	39.5

Table 1: Mean SI-SDRi (dB) of best performing models.



(a) Training the encoder-decoder parts using the proposed two-step approach.

(b) End-to-end separation network training using time-domain SI-SDR loss.

Fig. 2: Latent representations of a 1sec mixture and its constituent sources when training the same encoder architecture: a) individually using the proposed two-step approach (top) b) jointly with the TDCN separation module using SI-SDR loss on time-domain (bottom). We sort the basis indexes w.r.t. their energy and we raise the value of each cell to 0.1 for better visualization.

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.