






端到端语音分离与目标说话人抽取

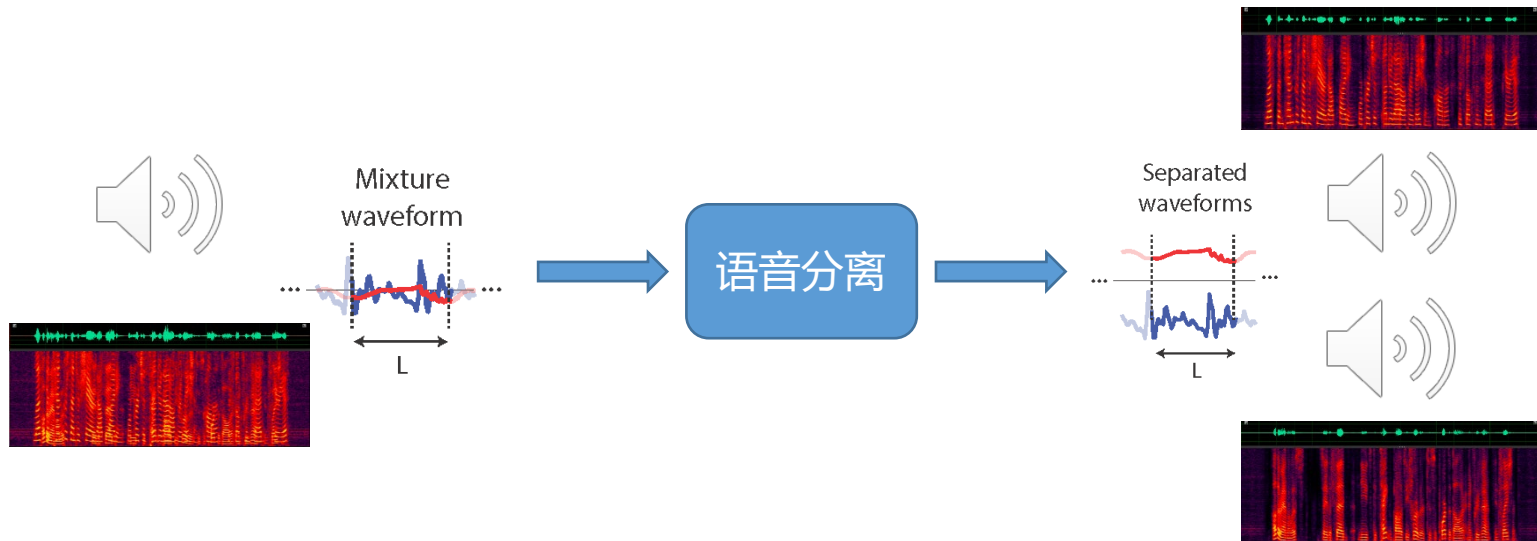
主讲人 宋辉

清华大学电子工程系 博士
滴滴AI Labs 语音技术部

- 
-  9.1 端到端语音分离的基本框架
 -  9.2 单通道语音分离和目标说话人抽取技术
 -  9.3 多通道语音分离技术
 -  9.4 实战

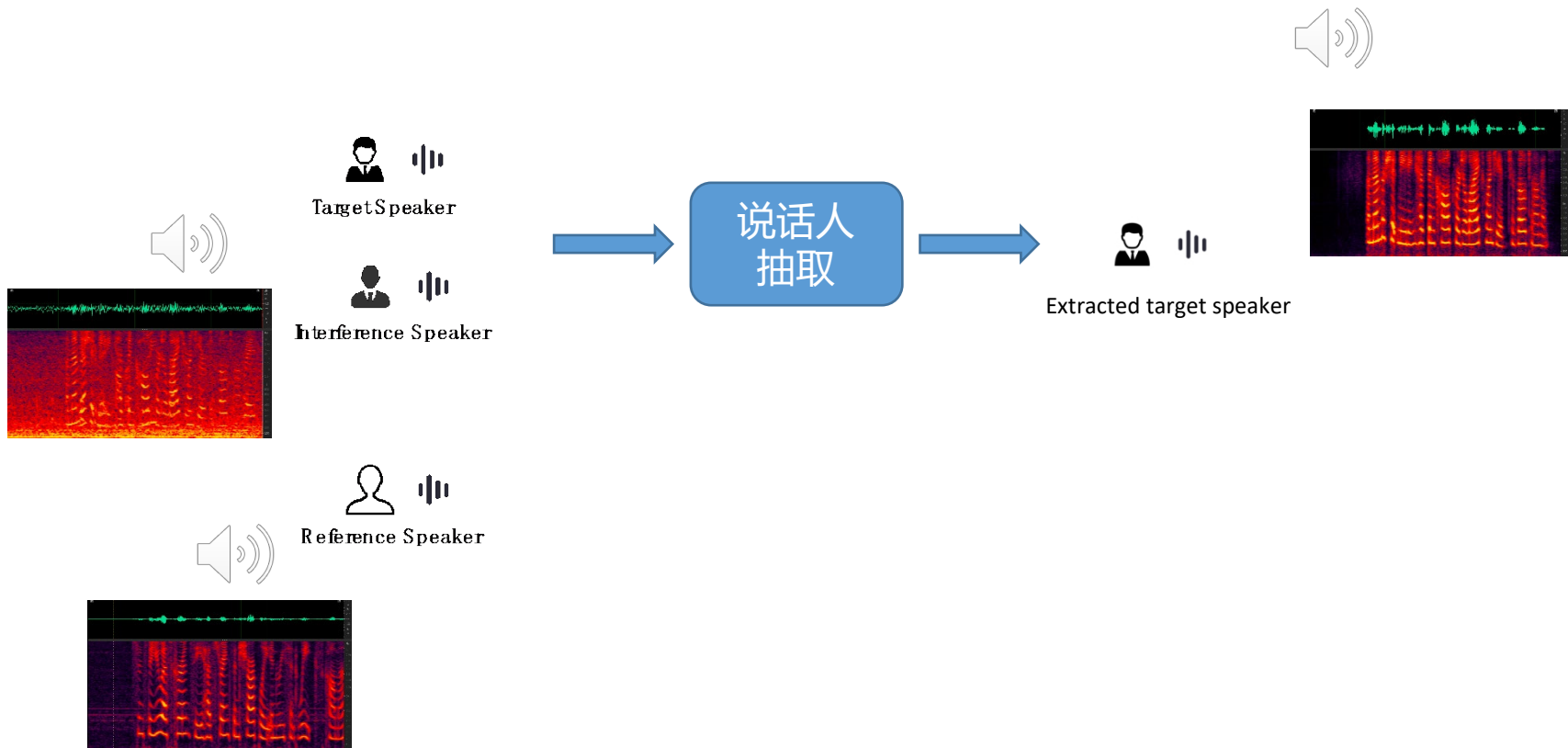


语音分离——示例





目标说话人抽取——示例





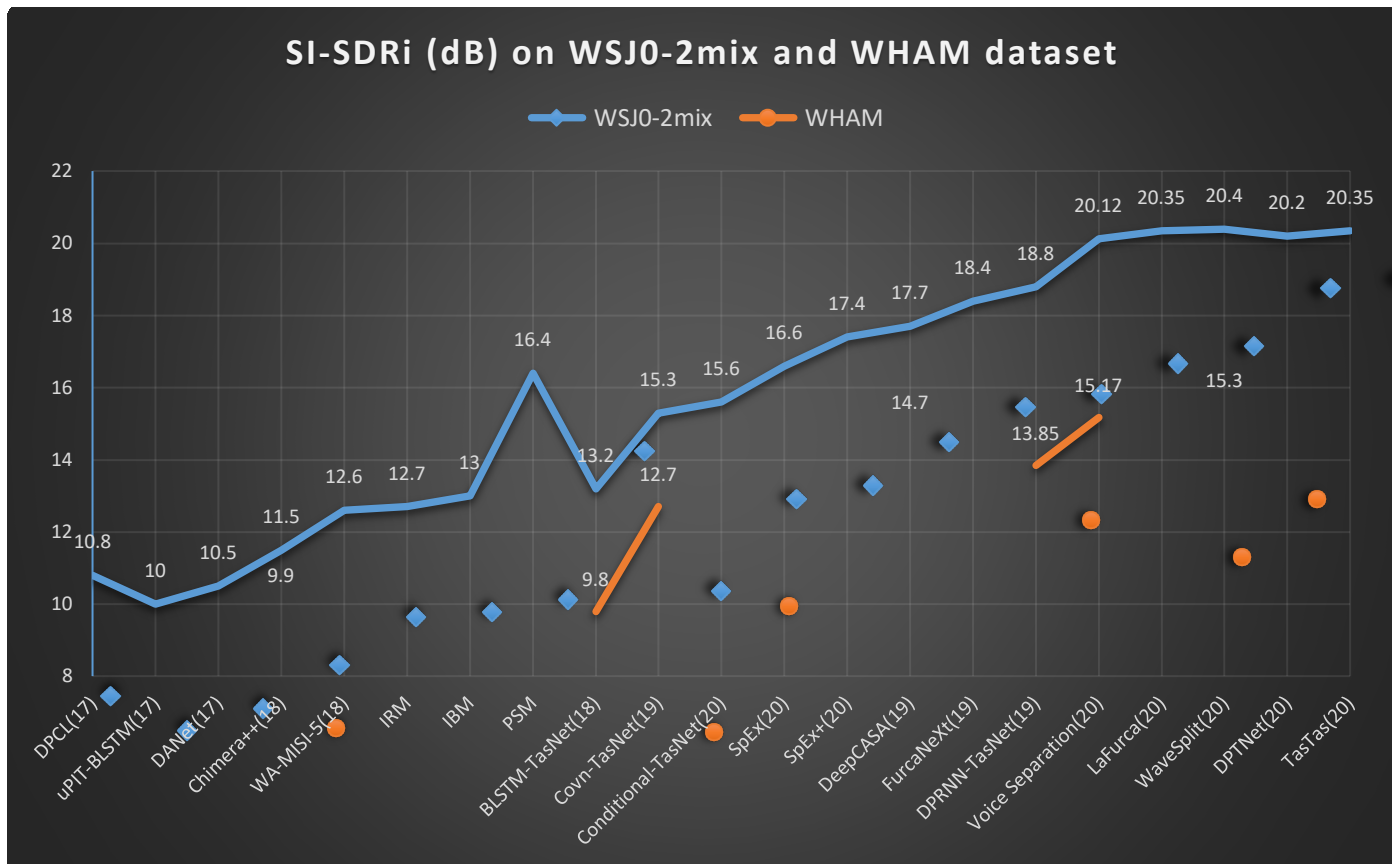
语音分离 & 目标说话人抽取——SOTA

时域 Vs 频域

噪声数据集下的结果
还不够完备

针对特定的测试集

分离 Vs 抽取





9.1 端到端语音分离的基本框架

Encoder-Separator-Decoder 框架:

Encoder:

-- 将输入信号从时域变换到另一个域 (domain) 或潜空间 (latent space) 中, 在潜空间中完成语音分离。

Separator (+ Extractor):

-- 在潜空间中, 为每个独立声源估计mask;
-- 通过元素级别相乘, 得到每个独立声源在潜空间中的估计。

Decoder:

-- 将分离后的各个源信号反变换回到时域。

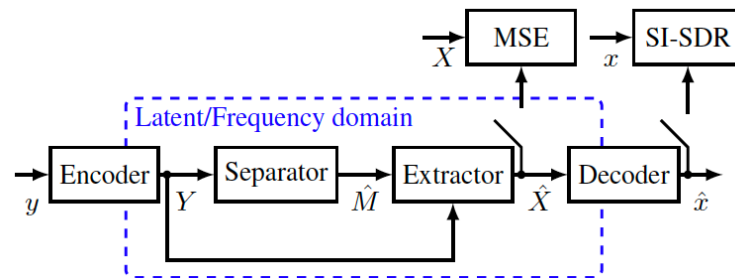


Fig. 1. Generic view of source separation system ^[1]



9.1 端到端语音分离的基本框架

时域观测信号^[2]:

$$x(t) = \sum_{i=1}^C s_i(t) + n(t), \quad (9.1)$$

Encoder变换:

$$\mathbf{X}(k, n) = \sum_{t=0}^{L-1} x(t + kH) u_n(t), \quad n \in \{0, \dots, N-1\}, \quad (9.2)$$

其中 k 为 frame index, L 为分析滤波器的长度, H 为 hop size, N 为分析滤波器个数。

Separator:

$$\mathcal{MN}(\mathcal{G}(\mathbf{X})) = [\mathbf{M}_1, \dots, \mathbf{M}_C]. \quad (9.3)$$

Extractor:

$$\mathbf{Y}_i = \mathcal{G}(\mathbf{X}) \odot \mathbf{M}_i, \quad i \in \{1, \dots, C\}, \quad (9.4)$$

Decoder变换:

$$\hat{s}_i(t) = \sum_{k=0}^{K-1} \sum_{n=0}^{N-1} \mathbf{Y}_i(k, n) v_n(t - kH). \quad (9.5)$$



9.1 端到端语音分离的基本框架

举例——handcrafted transformation (如STFT) :

$$\begin{aligned}u_n(t) &= h_a(t)e^{-2j\pi n/N} \\v_n(t) &= h_s(t)e^{2j\pi n/N},\end{aligned}\tag{9.6}$$

举例——learned transformation:

- 滤波器 $\{u_n(t)\}$ 和 $\{v_n(t)\}$ 的系数由网络自己学习, 它们既可以与separator联合学习, 也可以分别学习。
- *1-D CNN* 是将时域信号变换到潜空间中的一个典型应用。



基于time-frequency masking的分离方法

优点:

- 与经典的信号处理算法（如频域波束形成）更好的相容；
- mask具有较强的可解释性（interpretability）。

缺点:

- STFT是一种通用的信号变换工具，对于语音分离这一特定任务而言未必是最优的；
- 准确的相位重建比较困难；
- 需要较高的频率分辨率，所以延迟较大，不太适用于对实时性要求较高的场景。

Loss:

- 通常是MSE

举例:

- Deep Clustering ^[3], Deep attractor network ^[4], u-PIT ^[5], Deep CASA ^[6], Voice filter ^[7], SpeakerBeam ^[8], SBF-MTSAL-Concat ^[9], ...



基于时域的分离方法

优点:

- 利用数据驱动的思想, 让网络自己学习声音信号的声学特征表征, 取代STFT;
- 无需显式地处理相位重建的问题;
- 短延时 (如: 2ms in Conv-TasNet, 2 samples in DPRNN-TasNet)

缺点:

- mask的可解释性较弱;
- 不易与经典的信号处理算法相容;

Loss:

- SI-SDR

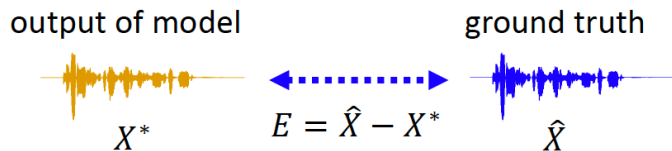
举例:

- TasNet^[10], Conv-TasNet^[11], Conditional-TasNet^[12], DPRNN-TasNet^[13], SpEx^[14], SpEx+^[15], Voice Separation^[16], LaFurca^[17], Wavesplit^[18], DPTNet^[19], TasTas^[20], ...

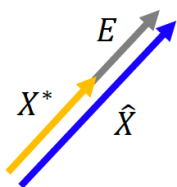


Scale invariant – signal-to-distortion ratio (SI-SDR)

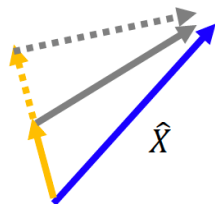
Signal-to-noise ratio (SNR):



$$SNR = 10 \log_{10} \frac{\|\hat{X}\|^2}{\|E\|^2}$$

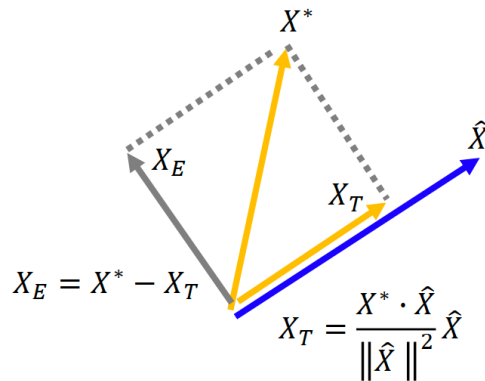


badcase 1



badcase 2

SI-SDR:



$$SISDR = 10 \log_{10} \frac{\|X_T\|^2}{\|X_E\|^2}$$

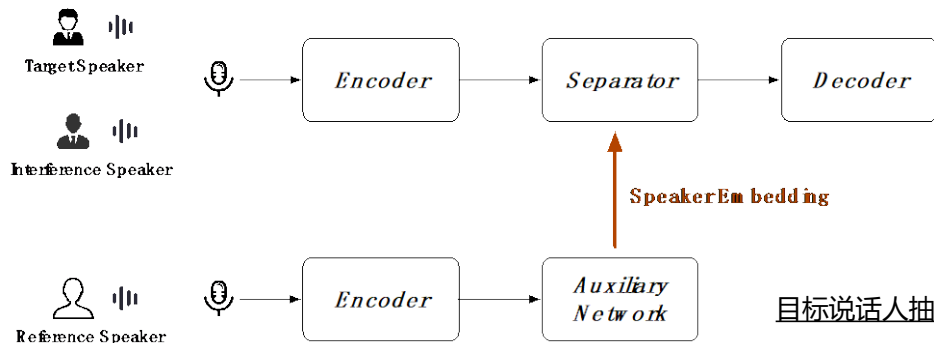
摘自李宏毅老师《speech separation》



目标说话人抽取

所谓目标说话人抽取 (Target Speaker Extraction), 是指从一段混合语音当中抽取出**特定说话人**语音的分离技术。

- “**说话人相关**” 的分离任务;
- 不需要关注**输出维度问题**;
- 不需要关注**排序问题**;
- 输出一路信号, 无需在多路信号中选择目标信号;
- 需要 “**参考信息**” 作为辅助。



目标说话人抽取技术框架

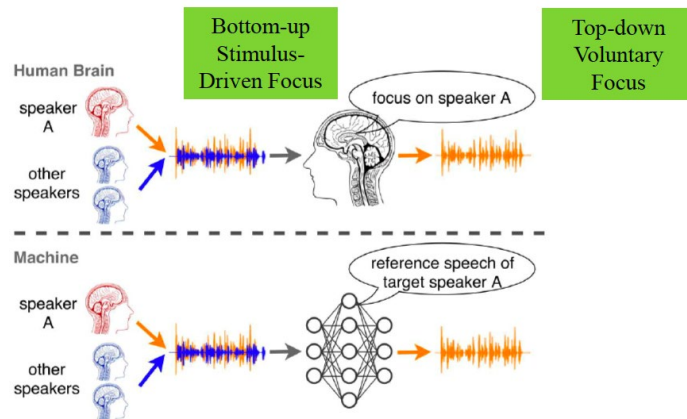
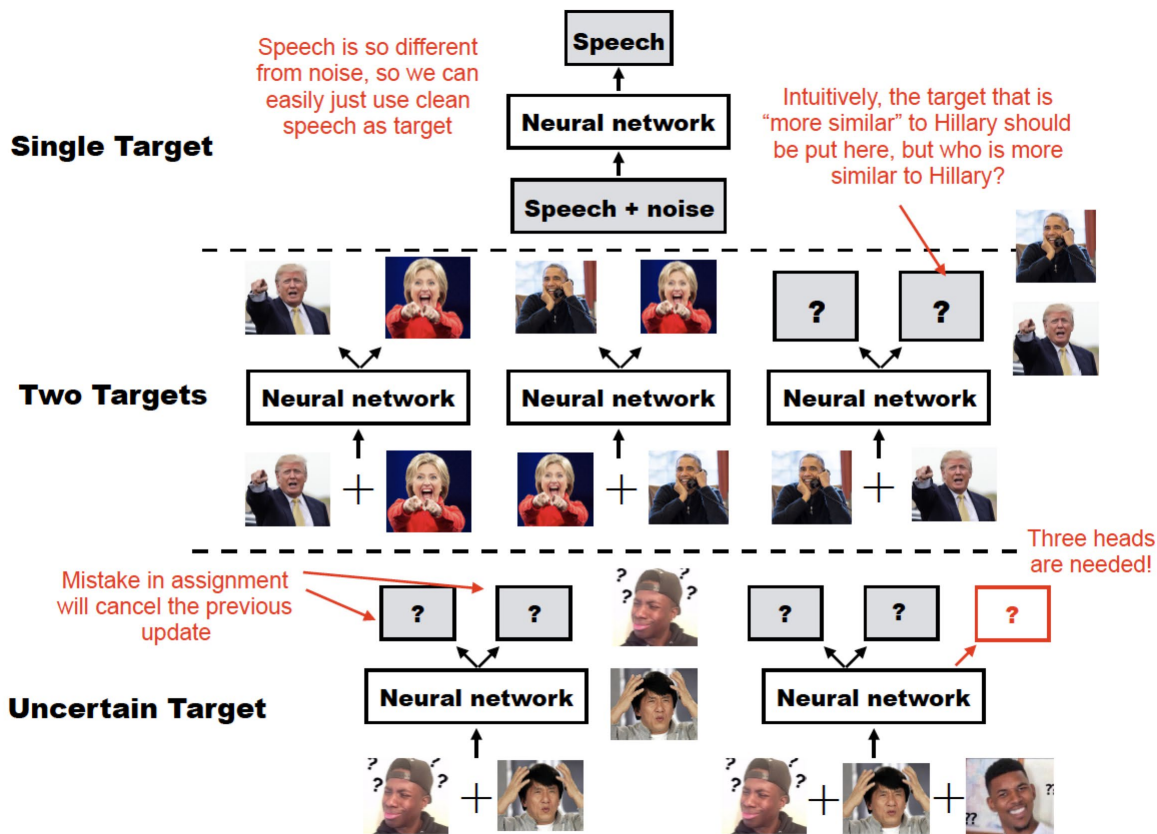


Fig. 2. Selective auditory attention ^[14]



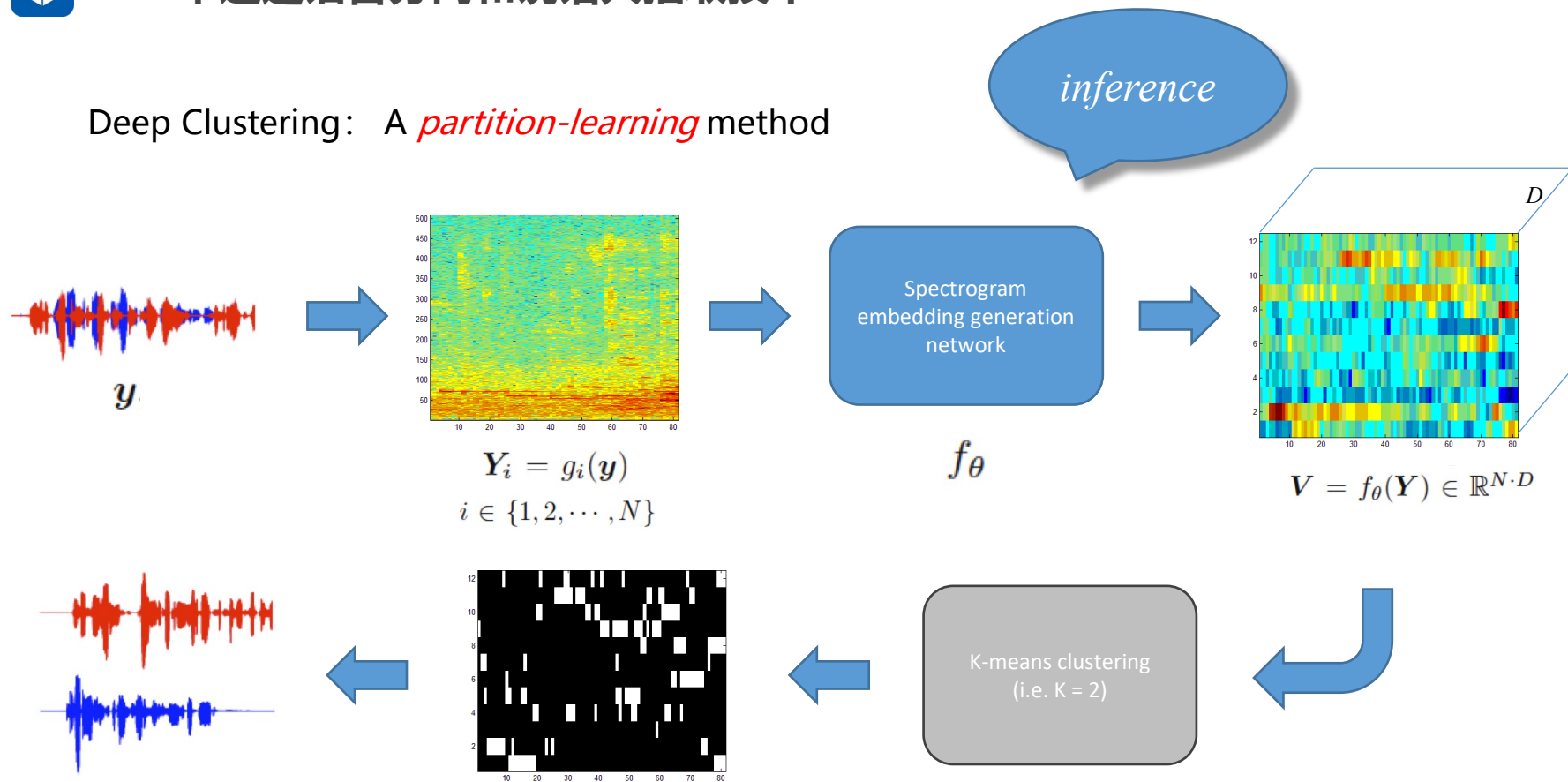
9.2 单通道语音分离和说话人抽取技术——排序问题和输出维度问题





9.2 单通道语音分离和说话人抽取技术——DPCL

Deep Clustering: A *partition-learning* method



9.2 单通道语音分离和说话人抽取技术——DPCL

Deep Clustering: A *partition-learning* method

Partition-based training:

-- 寻找能够准确聚类的embedding的生成方式。

-- 损失函数:

$$C(\theta) = \|VV^T - YY^T\|_W^2$$

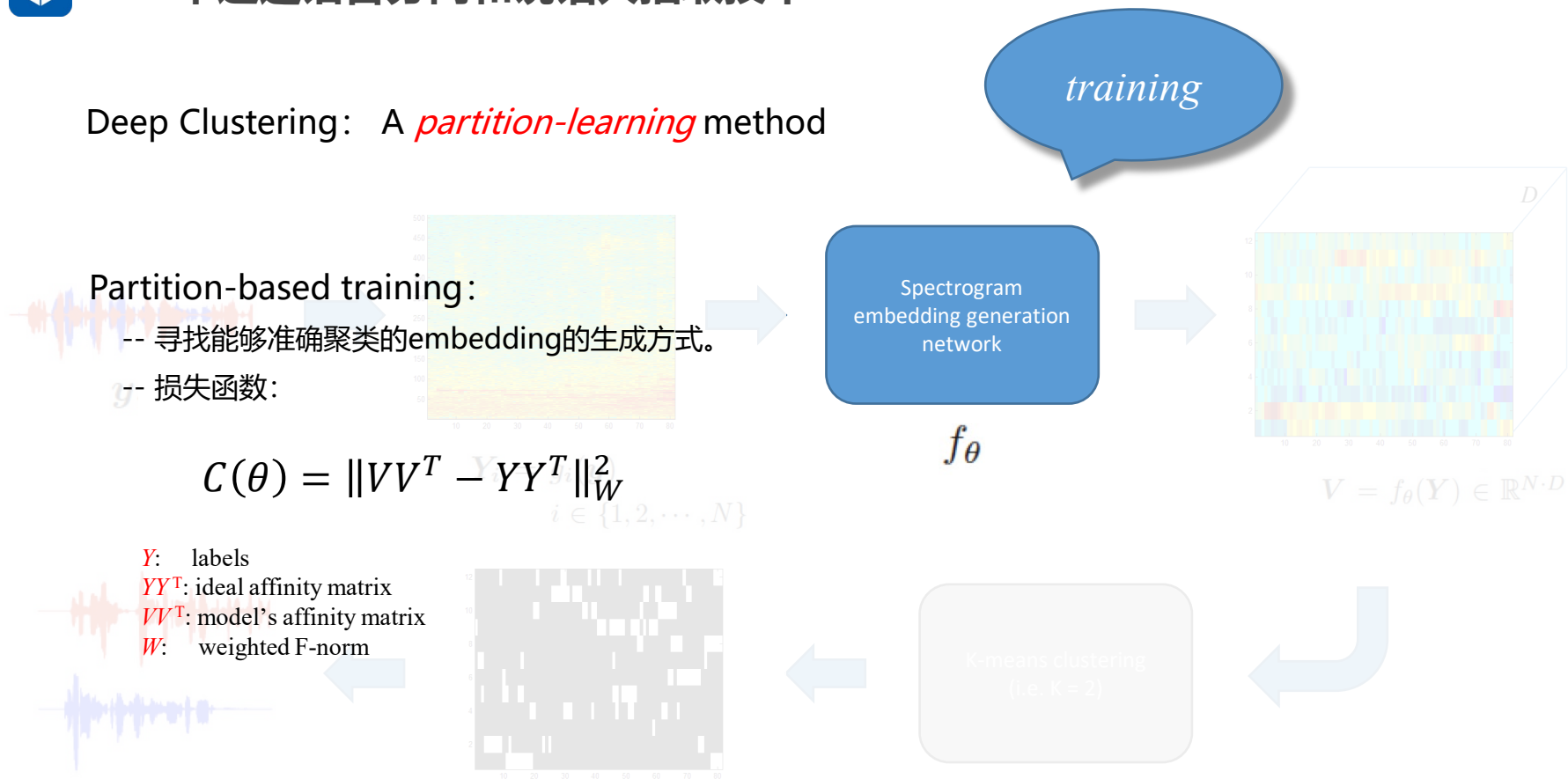
$i \in \{1, 2, \dots, N\}$

Y : labels

YY^T : ideal affinity matrix

VV^T : model's affinity matrix

W : weighted F-norm





9.2 单通道语音分离和说话人抽取技术——Permutation Invariant Training, PIT^[5]

生成mask的过程需要遍历全部的排序：

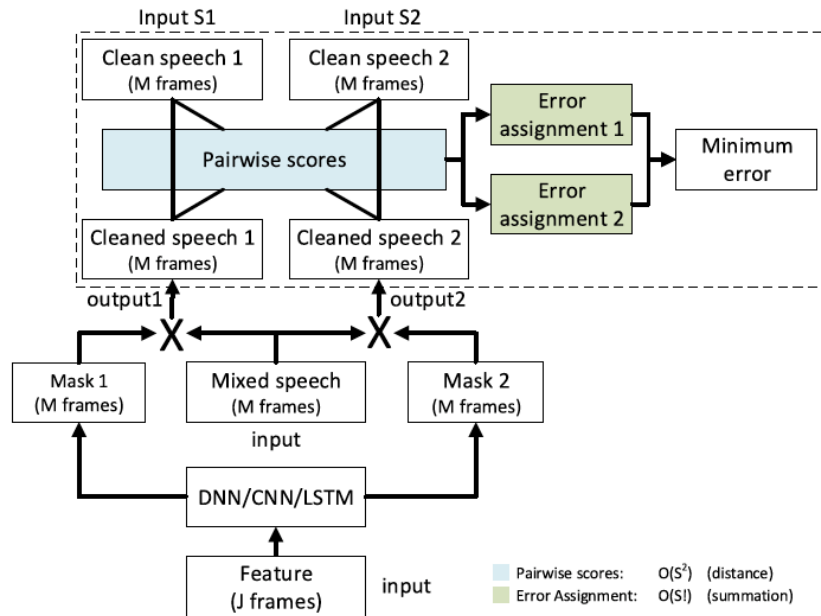
-- 给定一个分离网络，为混合语音中的每个声源生成对应的mask；

-- 穷举所有可能的排序 ($N!$) ；

-- label assignment: 使用loss最小的排序，更新分离网络；

-- 重复上述过程直至收敛；

-- 可以解决排序问题，但无法解决输出维度问题。



The two-talker speech separation model with permutation invariant training.



9.2 单通道语音分离和说话人抽取技术——Deep CASA^[6]

Computational Auditory Scene Analysis:

-- “分而治之”

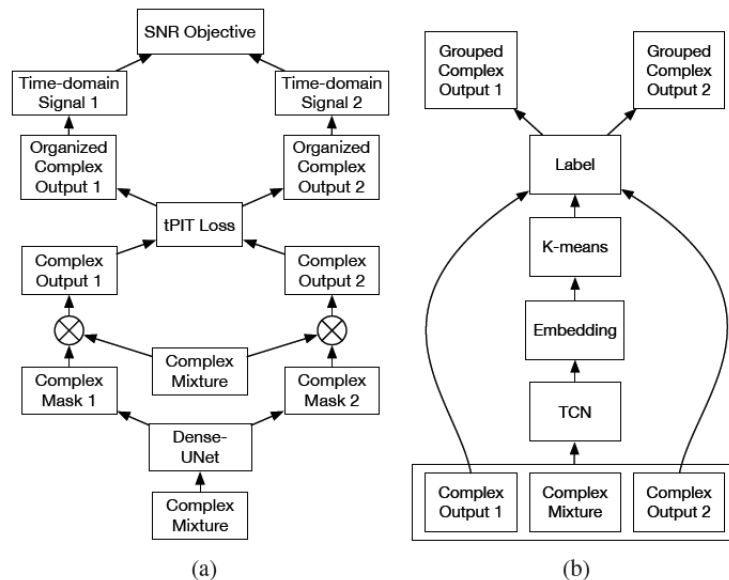
-- simultaneous grouping

利用PIT实现帧级别分离，即分离出同一帧里分别属于不同声源的谱分量。

-- sequential grouping

利用聚类网络实现每个独立说话人声源的追踪。

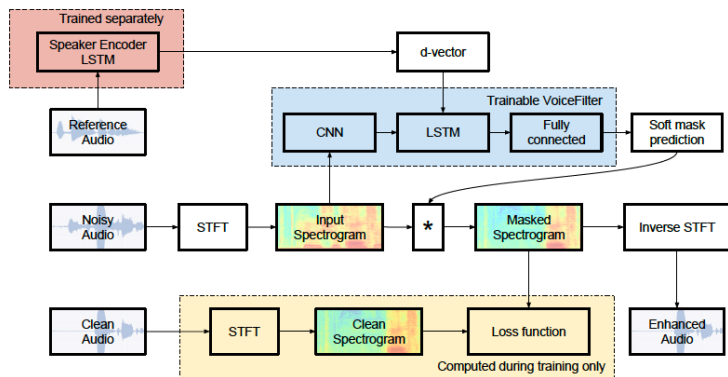
训练规则类似DPCL，属于同一说话人的embedding距离尽可能接近，属于不同说话人的embedding距离尽可能远离。



Diagrams of (a) the simultaneous grouping stage and (b) the sequential grouping stage in deep CASA.



9.2 单通道语音分离和说话人抽取技术——Voice filter



VoiceFilter^[7]

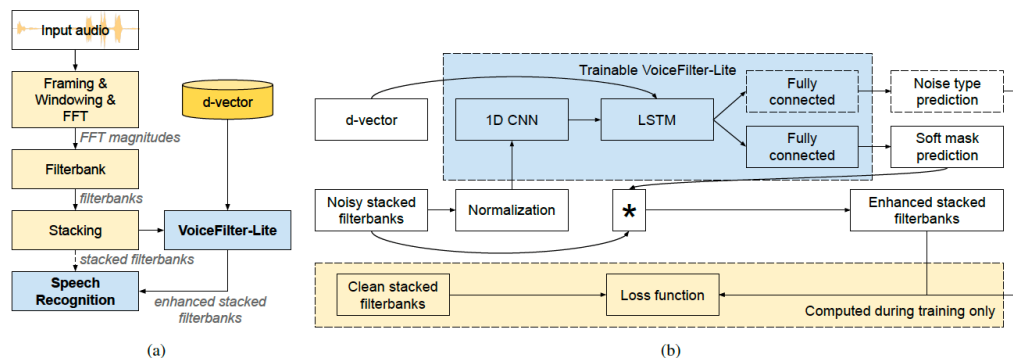


Figure 1: VoiceFilter-Lite architecture, assuming using stacked filterbank energies as inputs and outputs. (a) Integration with ASR. The dashed arrow indicates the original connection without VoiceFilter-Lite. (b) Neural network topology of the VoiceFilter-Lite model.

VoiceFilter-Lite^[21]:

- 面向ASR任务，无需重建信号波形；
- 输入：声学特征，*e.g. stacked filterbank energies*；
- 输出：增强特征；
- 避免over-suppression问题：

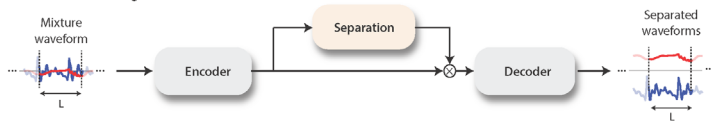
$$L_{\text{asym}} = \sum_t \sum_f \left(g_{\text{asym}}(S_{\text{cIn}}(t, f) - S_{\text{enh}}(t, f), \alpha) \right)^2.$$



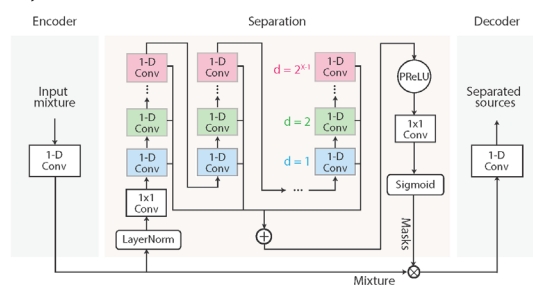
9.2 单通道语音分离和说话人抽取技术——TasNet

Conv-TasNet^[11]:

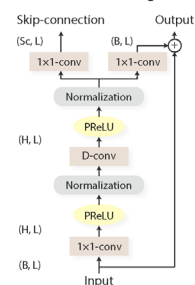
A. TasNet block diagram



B. System flowchart



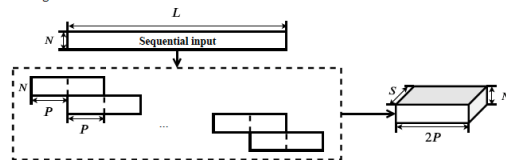
C. 1-D Conv block design



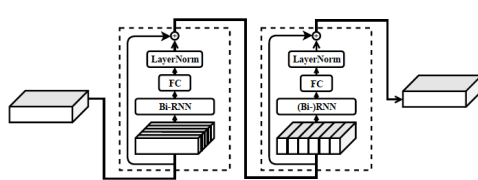
Dual-Path RNN-TasNet^[13]:

-- utterance-level processing

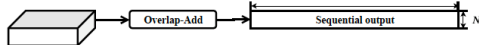
A. Segmentation



B. DPRNN block



C. Overlap-Add



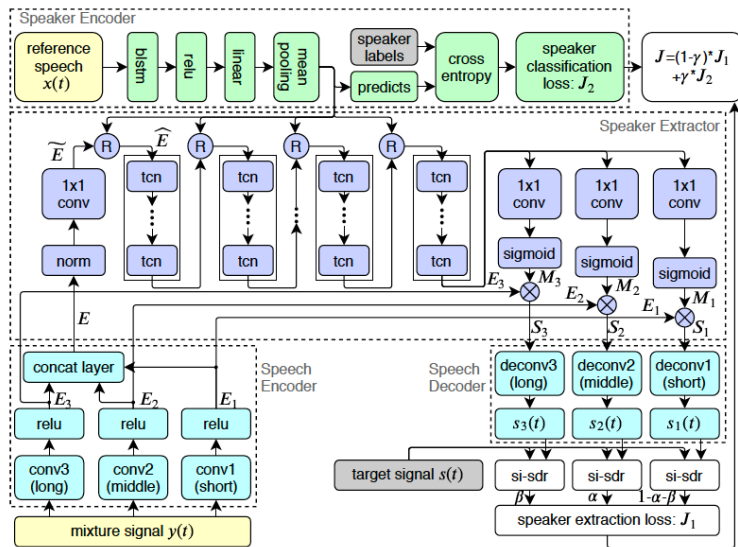
Multi-Path RNN-TasNet^[22]:

-- inter-utterance-level processing

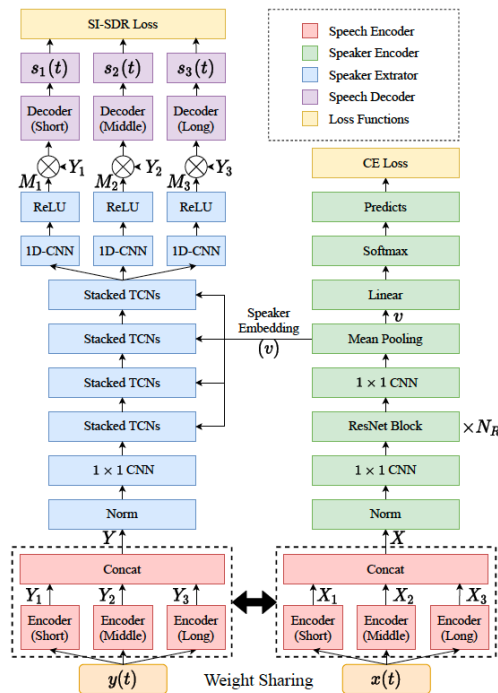


9.2 单通道语音分离和说话人抽取技术——SpEx & SpEx+

SpEx [14]:

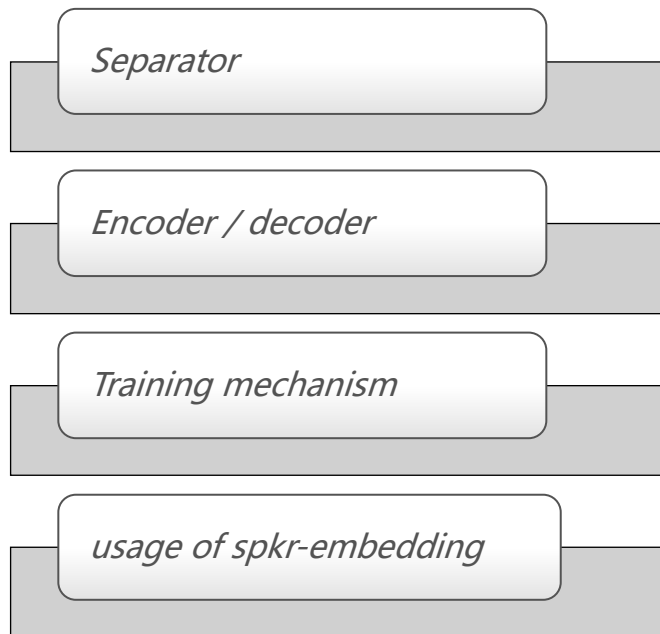


SpEx+ [15]:



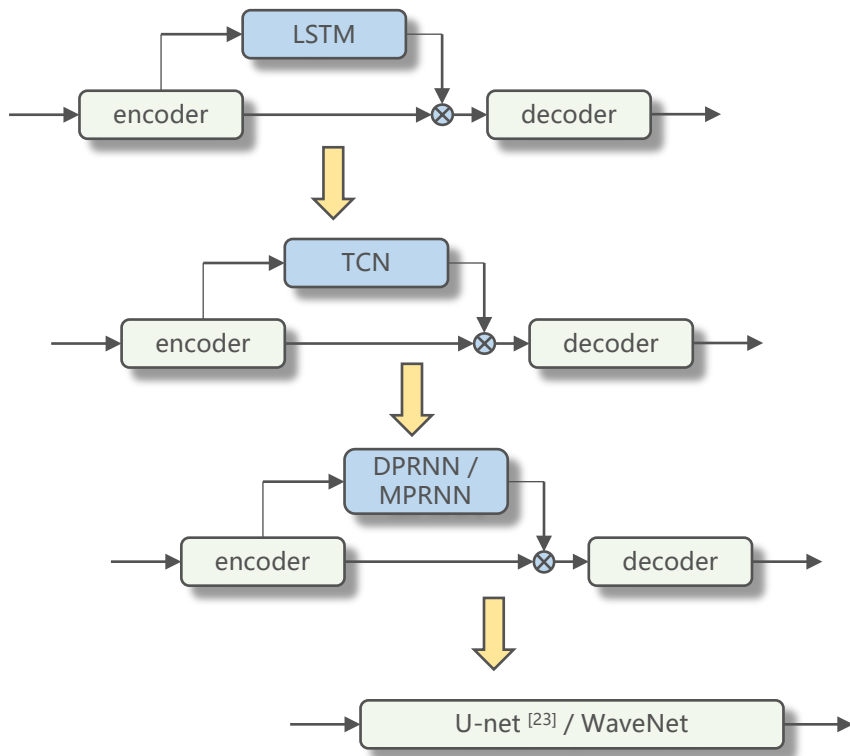
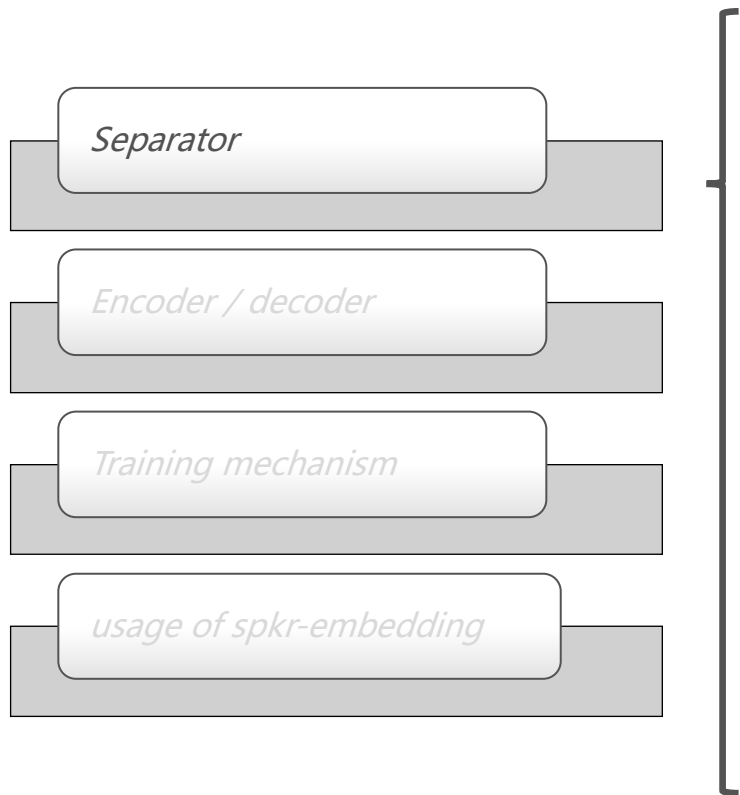


9.2 单通道语音分离和说话人抽取技术——技术脉络



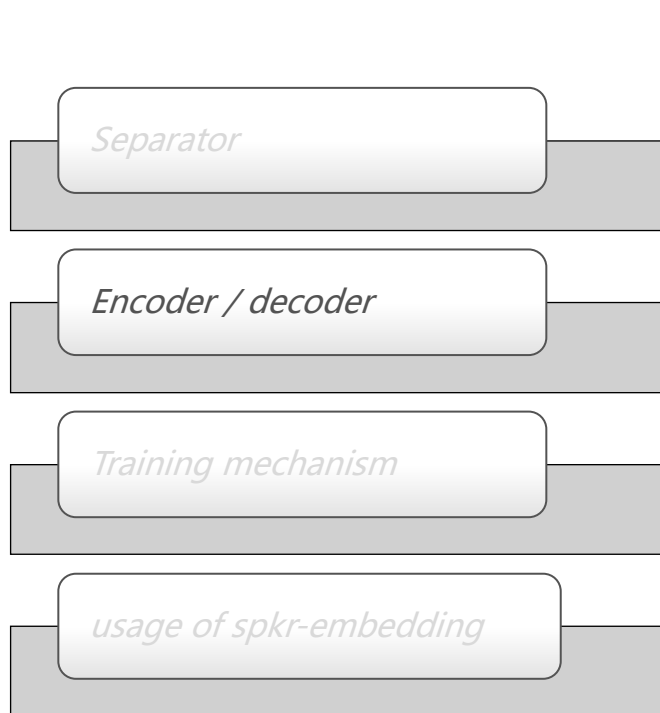


9.2 单通道语音分离和说话人抽取技术——分离器





9.2 单通道语音分离和说话人抽取技术——编码器和解码器



- The encoder / decoder fall into 3 categories ^{[2][23]}:

$$\mathbf{X}(k, n) = \sum_{t=0}^{L-1} x(t + kH) u_n(t), \quad n \in \{0, \dots, N-1\},$$

- **固定滤波器** *Fixed filters* : STFT, mel filterbank, gammatone filterbank, ...
- **参数化滤波器** *Parameterized filters* : a family of filters whose parameters are learned with the network.
- **自由滤波器** *Free filters* : all weights are jointly or separately learned.

Table 3: Comparison of different encoder and decoder combination using $L_w = 4$ ms and $L_s = 2$ ms on the test set of the WSJ0-2mix database.

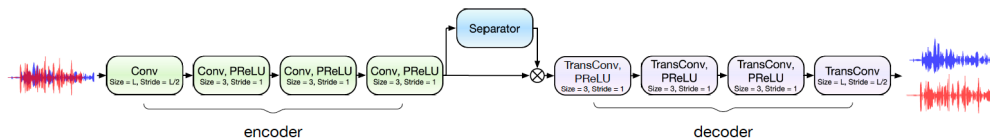
Loss-Fn	Encoder	Decoder	si-SDR dB	SDR dB	WER %
$\mathcal{L}^{\text{SI-SDR}}$	learned	learned	14.4	14.7	21.71
$\mathcal{L}^{\text{SI-SDR}}$	STFT	learned	13.9	14.3	21.92
$\mathcal{L}^{\text{SI-SDR}}$	learned	ISTFT	14.1	14.5	21.87
$\mathcal{L}^{\text{SI-SDR}}$	STFT	ISTFT	12.4	12.8	24.69



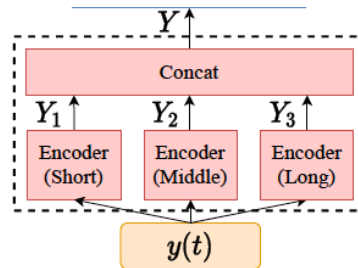
9.2 单通道语音分离和说话人抽取技术——编码器和解码器



- Deep encoder / decoder [25]:

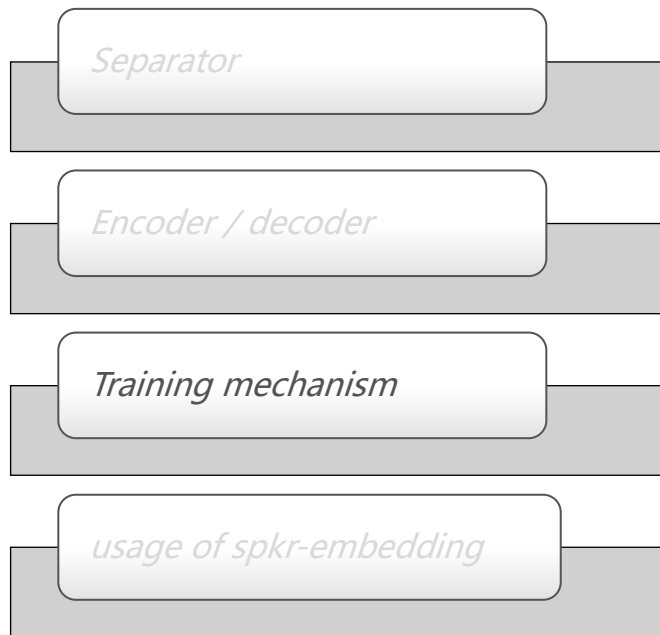


- Multi-scale encoder :
 - 平均可带来 **0.3dB** 的SI-SDR提升。

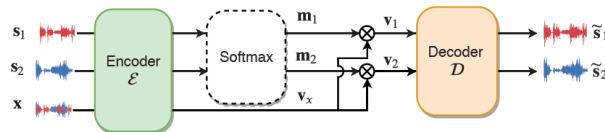




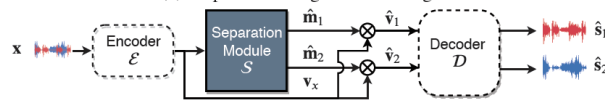
9.2 单通道语音分离和说话人抽取技术——训练机制



- Two-step training [26]:



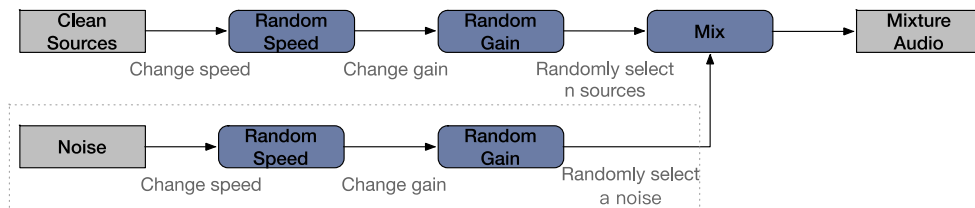
(a) Step 1: Learning the latent targets.



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

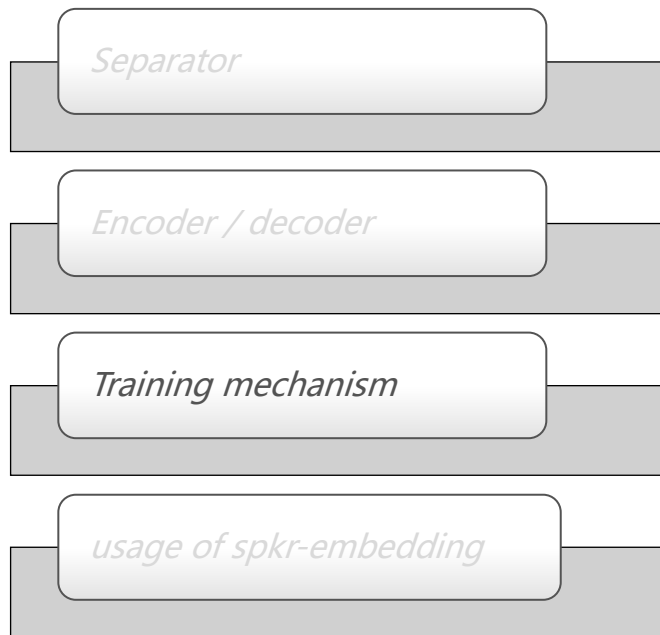
Fig. 1: Training a separation network in two independent steps. For each step, the non-trainable parts are represented with a dashed line.

- Dynamic mixing for data augmentation :





9.2 单通道语音分离和说话人抽取技术——训练机制



- Mixup-breakdown training (MBT) ^[27]:

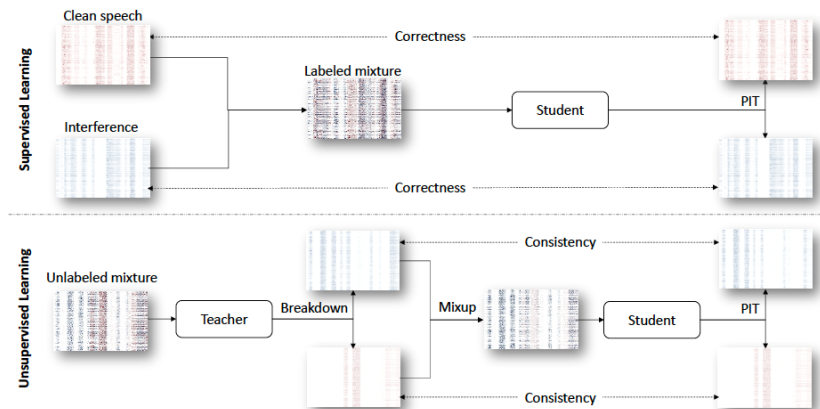


Fig. 1: Algorithmic flow of the Mixup-Breakdown training divided into the supervised and unsupervised learning procedures

$$\text{Mix}_{\lambda}(\mathbf{a}, \mathbf{b}) \triangleq \lambda \cdot \mathbf{a} + (1 - \lambda) \cdot \mathbf{b}$$

$$\text{Break}_{\lambda}(\mathbf{a}, \mathbf{b}) \triangleq (\lambda \cdot \mathbf{a}, (1 - \lambda) \cdot \mathbf{b})$$

$$\mathbf{f}_{\theta_S}(\text{Mix}_{\lambda}(\mathbf{f}_{\theta_T}(\mathbf{x}_j))) \approx \text{Break}_{\lambda}(\mathbf{f}_{\theta_T}(\mathbf{x}_j))$$



9.2 单通道语音分离和说话人抽取技术——训练机制

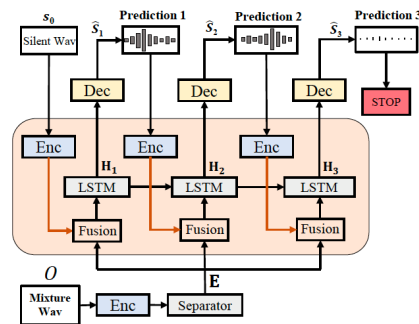
Separator

Encoder / decoder

Training mechanism

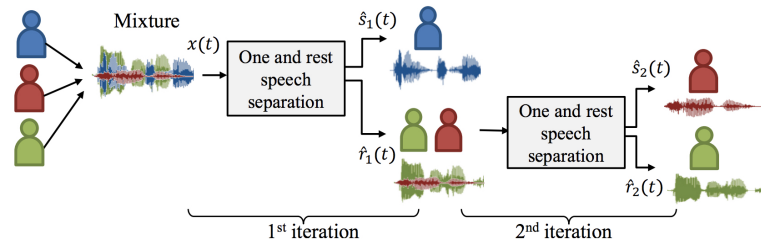
usage of spkr-embedding

- Conditional chain model [12][28]:
 - 适用于说话人个数未知的场景。



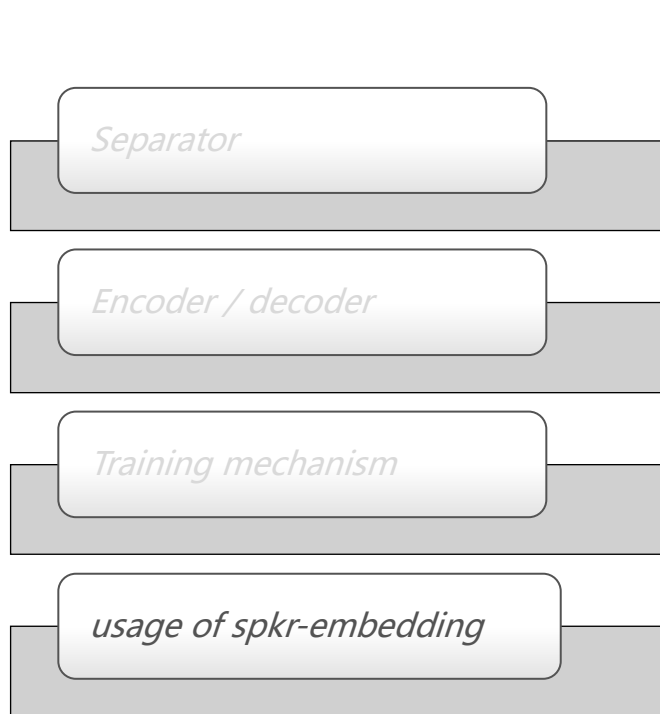
(a) Conditional Chain Model for speech separation

- One-and-rest PIT, OR-PIT [33]:





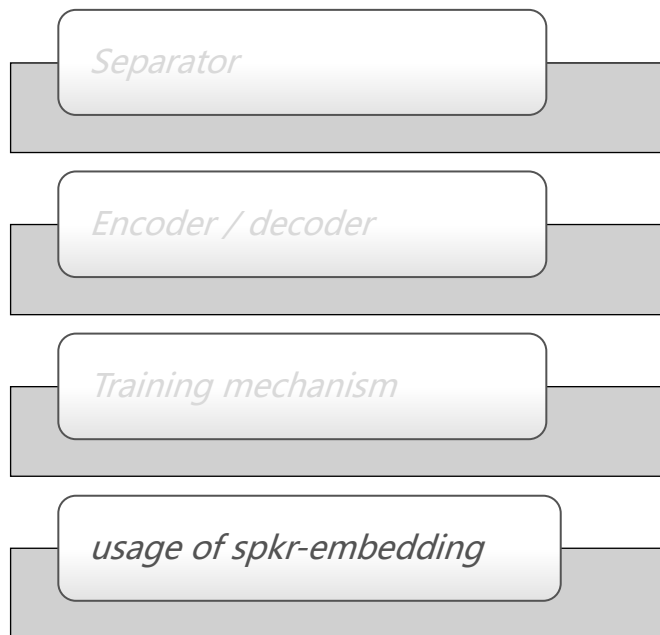
9.2 单通道语音分离和说话人抽取技术——如何用好spkr-embedding



- 无embedding:
 - Conv-TasNet, DPRNN-TasNet, ...
- 固定的embedding:
 - Fbank, x-vector, d-vector, ...
 - 例如: *Voice filter*
- 网络学习出的embedding:
 - 通常会有一个显式的 *speaker encoder* 与抽取器联合训练;
 - 通常会引入CE loss;
 - *speaker encoder* 与 *speech encoder* 的配合;
 - 例如: *SpEx*, *SpEx+*, *Wavesplit*, *Atss-Net*



9.2 单通道语音分离和说话人抽取技术——如何用好spkr-embedding



- Wavesplit [18]:

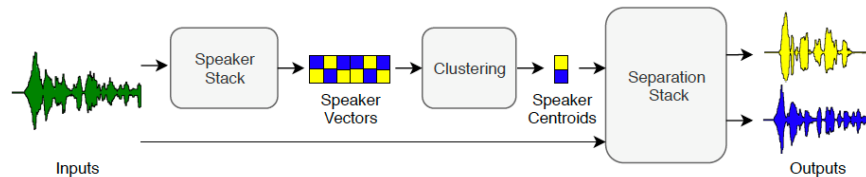


Figure 1: Wavesplit for 2-speaker separation. The speaker stack extracts speaker vectors at each timestep. The vectors are clustered and aggregated in speaker centroids. The separation stack ingests the centroids and the input signal to output two clean channels.

- Atss-Net [29]:

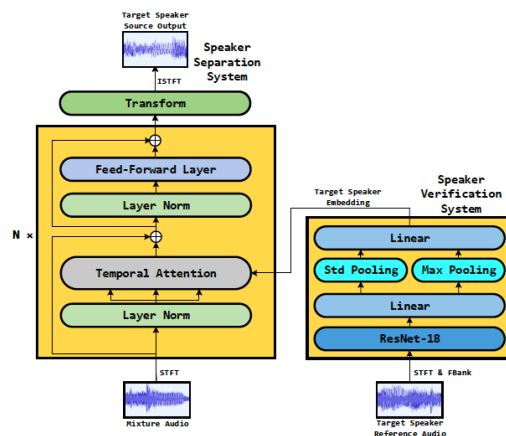


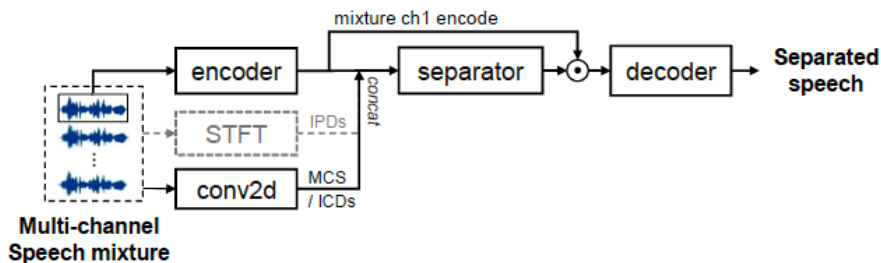
Figure 2: Model architecture of the Atss-Net.



9.3 多通道语音分离技术

Separation-wise methods:

- 利用多通道带来的空间信息帮助一个分离网络更好的工作;
- 使用固定类空间特征: IPD、ILD、...;
- 使用网络学习的空间特征^[30];





9.3 多通道语音分离技术

Separation-wise methods:

Beamforming-wise methods:

- 将分离网络看作一个前端模块，帮助Beamformer更好的工作；
- Beam-TasNet: 为频域波束形成（如MVDR）提供更可靠的统计量估计^[31]；

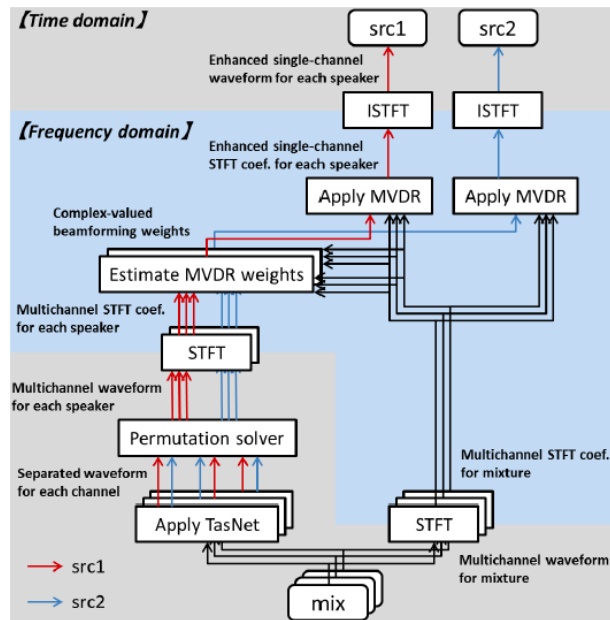


Fig. 1. Overall separation procedures in Beam-TasNet

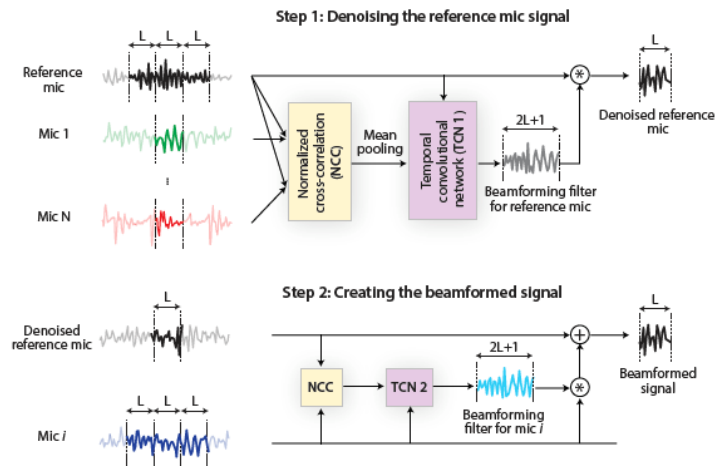


9.3 多通道语音分离技术

Separation-wise methods:

Beamforming-wise methods:

- 将分离网络看作一个前端模块，帮助Beamformer更好的工作；
- Beam-TasNet: 为频域波束形成（如MVDR）提供更可靠的统计量估计^[31]；
- FaSNet: 直接估计波束形成器中每个滤波器的系数^[32]。





9.3 多通道语音分离技术

Separation-wise methods:

Beamforming-wise methods:

Pipeline / end-to-end solution:

- unmixing, fix-beamformer, extraction (UFE) [34]
- end-to-end UFE (E2E-UFE)

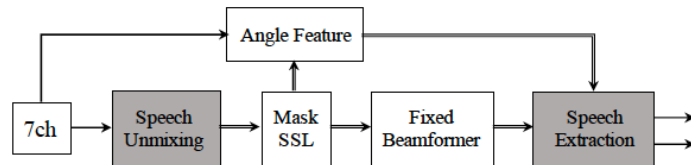
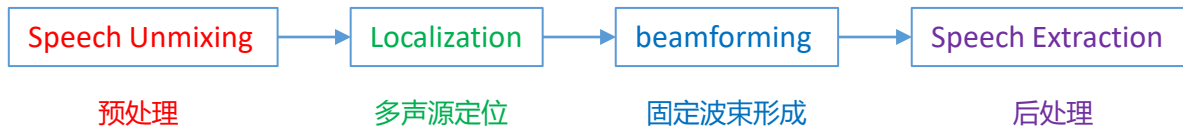


Figure 1: Overview of the UFE system. The grey block is an neural network trained independently.





本章回顾



9.1 端到端语音分离的基本框架



9.2 单通道语音分离和目标说话人抽取技术



9.3 多通道语音分离技术



- [1] Jens Heitkaemper, et al. "Demystifying TasNet: A dissecting approach," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6354-6358.
- [2] Manuel Pariente, et al. "Filterbank design for end-to-end speech separation," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6359-6363.
- [3] John.R. Hershey, et al. "Deep clustering: Discriminative embeddings for segmentation and separation," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 31-35.
- [4] Zhuo Chen, Yi Luo, Nima Mesgarani. "Deep attractor network for single-microphone speaker separation," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 246-250.
- [5] Morten Kolbaek, Dong Yu, Zheng-Hua Tan, Jesper 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] Yuzhou Liu, DeLiang Wang. "Deep CASA for talker-independent monaural speech separation," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6349-6353.
- [7] Quan Wang, et al. "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," arXiv preprint arXiv: 1810.04826, 2019.
- [8] Marc Delcroix, et al. "Single channel target speaker extraction and recognition with speaker beam," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5554-5558.
- [9] Chenglin Xu, Wei Rao, Eng Siong Chng, Haizhou Li. "Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6990-6994.
- [10] Yi Luo, Nima Mesgarani. "TasNet: time-domain audio separation network for real-time, single-channel speech separation," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.



- [11] Yi Luo, Nima 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.
- [12] Jing Shi, et al. "Sequence to multi-sequence learning via conditional chain mapping for mixture signals," arXiv preprint arXiv: 2006.14150, 2020.
- [13] Yi Luo, Zhuo Chen, Takuya 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.
- [14] Chenglin Xu, Wei Rao, Eng Siong Chng, Haizhou Li. "SpEx: Multi-scale time domain speaker extraction network," arXiv preprint arXiv: 2004.08326, 2020.
- [15] Meng Ge, Chenglin Xu, Longbiao Wang, Eng Siong Chng, Jianwu Dang, Haizhou Li. "SpEx+: A complete time domain speaker extraction network," arXiv preprint arXiv: 2005.04686, 2020.
- [16] Eliya Nachmani, Yossi Adi, Lior Wolf. "Voice separation with an unknown number of multiple speakers," arXiv preprint arXiv: 2003.01531, 2020.
- [17] Ziqiang Shi, Rujie Liu, Jiqiang Han. "LaFurca: Iterative multi-stage refined end-to-end monaural speech separation based on context-aware dual-path deep parallel inter-intra BiLSTM," arXiv preprint arXiv: 2001.08998, 2020.
- [18] Neil Zeghidour, David Grangier. "Wavesplit: End-to-end speech separation by speaker clustering," arXiv preprint arXiv: 2002.08933, 2020.
- [19] Jingjing Chen, Qirong Miao, Dong Liu. "Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation," in Interspeech 2020, pp. 2642 – 2646.
- [20] Ziqiang Shi, Rujie Liu, Jiqiang Han. "Speech Separation Based on Multi-Stage Elaborated Dual-Path Deep BiLSTM with Auxiliary Identity Loss," in Interspeech 2020, pp. 2682 – 2686.



参考文献

- [21] Quan Wang, et al. "VoiceFilter-Lite: Streaming Targeted Voice Separation for On-Device Speech Recognition," in Interspeech 2020, pp. 2677 – 2681.
- [22] Keisuke Kinoshita, et al. "Multi-path RNN for hierarchical modeling of long sequential data and its application to speaker stream separation," in Interspeech 2020, pp. 2652 – 2656.
- [23] Daniel Stoller, Sebastian Ewert, Simon Dixon. "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," arXiv preprint arXiv: 1806.03185.
- [24] David Ditter, Timo Gerkmann. "A multi-phase gammatone filterbank for speech separation via tasnet," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 36-40.
- [25] Berkan Kadioglu, et al. "An Empirical study of Conv-TasNet," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7259-7263.
- [26] Efthymios Tzinis, et al. "Two-step sound source separation: training on learned latent targets," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 31-35.
- [27] Max W.Y. Lam, Jun Wang, Dan Su, Dong Yu. "Mixup-breakdown: A consistency training method for improving generalization of speech separation models," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6369-6373.
- [28] Jing Shi, Jiaming Xu, Yusuke Fujita, Shinji Watanabe, Bo Xu. "Speaker-conditional chain model for speech separation and extraction," in Interspeech 2020. pp. 2707 – 2711.
- [29] Tingle Li, Qingjian Lin, Yuanyuan Bao, Ming Li. "Atss-Net: Target speaker separation via attention-based neural network," in Interspeech 2020, pp. 1411 – 1415.
- [30] Rongzhi Gu, Shi-xiong Zhang, Lianwu Chen, Yong Xu, Meng Yu, Dan Su, Yuexian Zou, Dong Yu. "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7319-7323.



参考文献

- [31] Tsubasa Ochiai, et al. "Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6379-6383.
- [32] Yi Luo, Cong Han, Nima Mesgarani. "FasNet: Low-latency adaptive beamforming for multi-microphone audio processing," ASRU 2019, pp. 260-267.
- [33] Naoya Takahashi, et al. "Recursive speech separation for unknown number of speakers," in Interspeech 2019. pp. 1348 – 1352.
- [34] Jian Wu, Zhuo Chen, Jinyu Li, Takuya Yoshioka, Zhili Tan, Ed Lin, Yi Luo, Lei Xie. "An End-to-end Architecture of Online Multi-channel Speech Separation," in Interspeech 2020. pp. 81 – 85.

感谢聆听！

Thanks for Listening

