Spectron: Target Speaker Extraction using Conditional Transformers

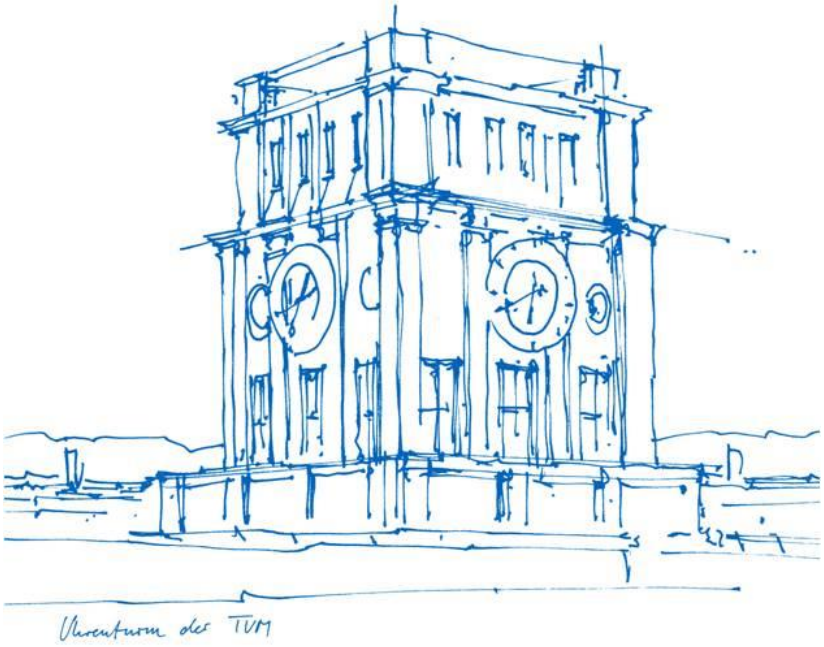
Tathagata Bandyopadhyay

Technische Universität München

Department of Informatics

Visual Computing Lab

Munich, 27 May 2022



Introduction

The Cocktail Party Problem

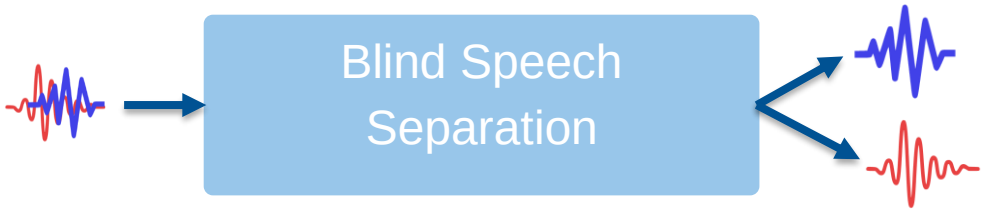
The problem:

- Many people talking simultaneously
- You want to listen to just one of them

Conventional Solutions:

- Multi-channel blind source separation
- **Single-Channel** blind source separation

Source: elixirofknowledge.com



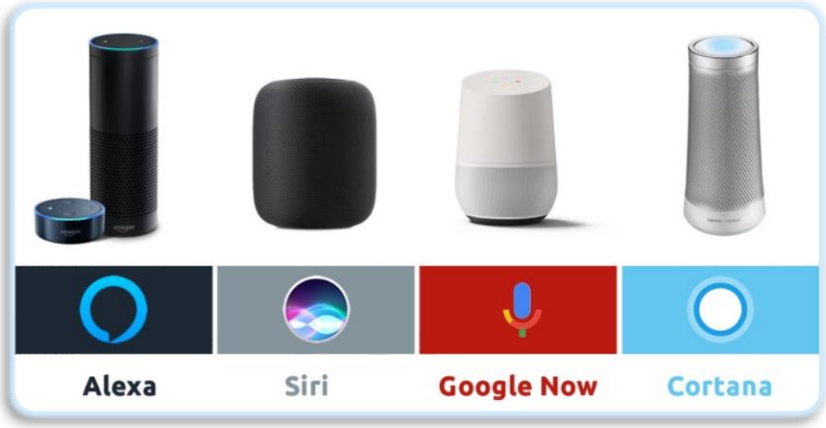
Informed is often better than Blind

Blind Source Separation:

- number of speakers need to be known
- Output channel assignment ambiguity
- Permutation Invariant Loss is not scalable

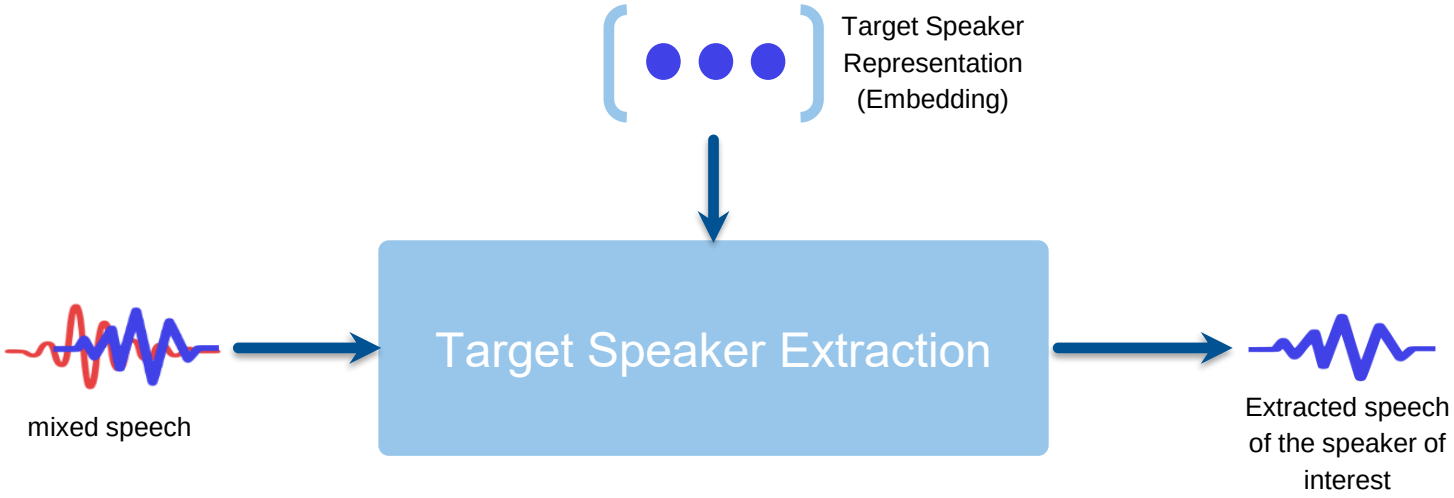
Often we know whom to look for:

- Personal Assistants
- Voice Commands
- Target Speech Recognition



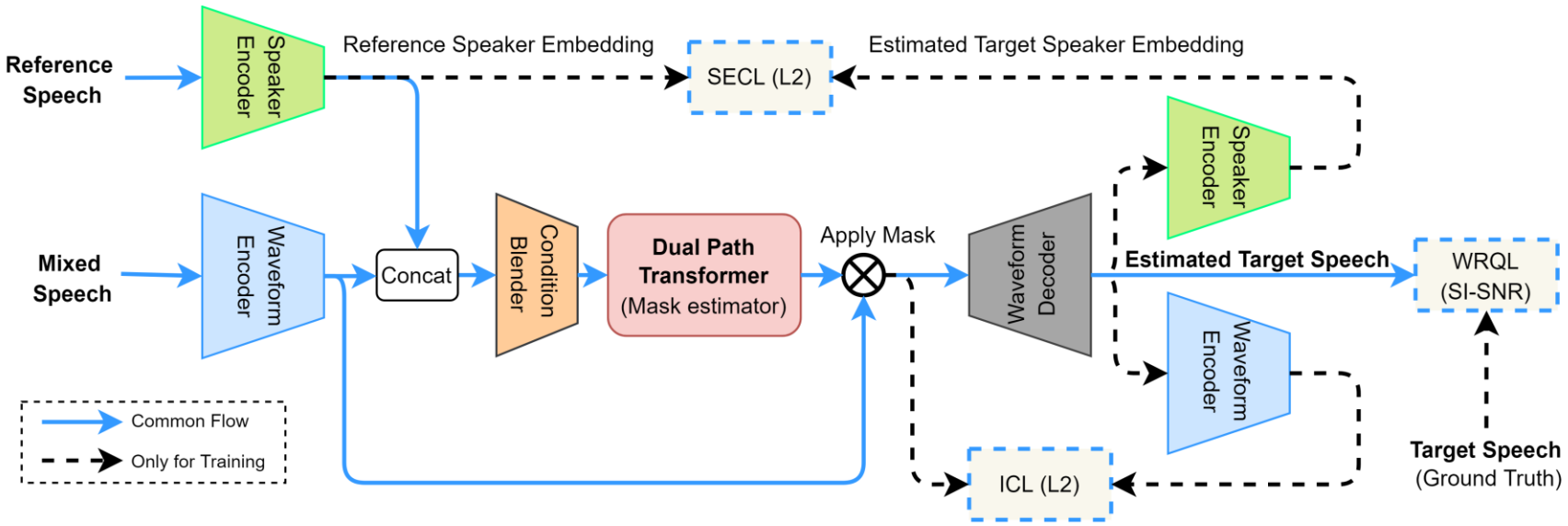
Source: medium.com

Target Speaker Extraction

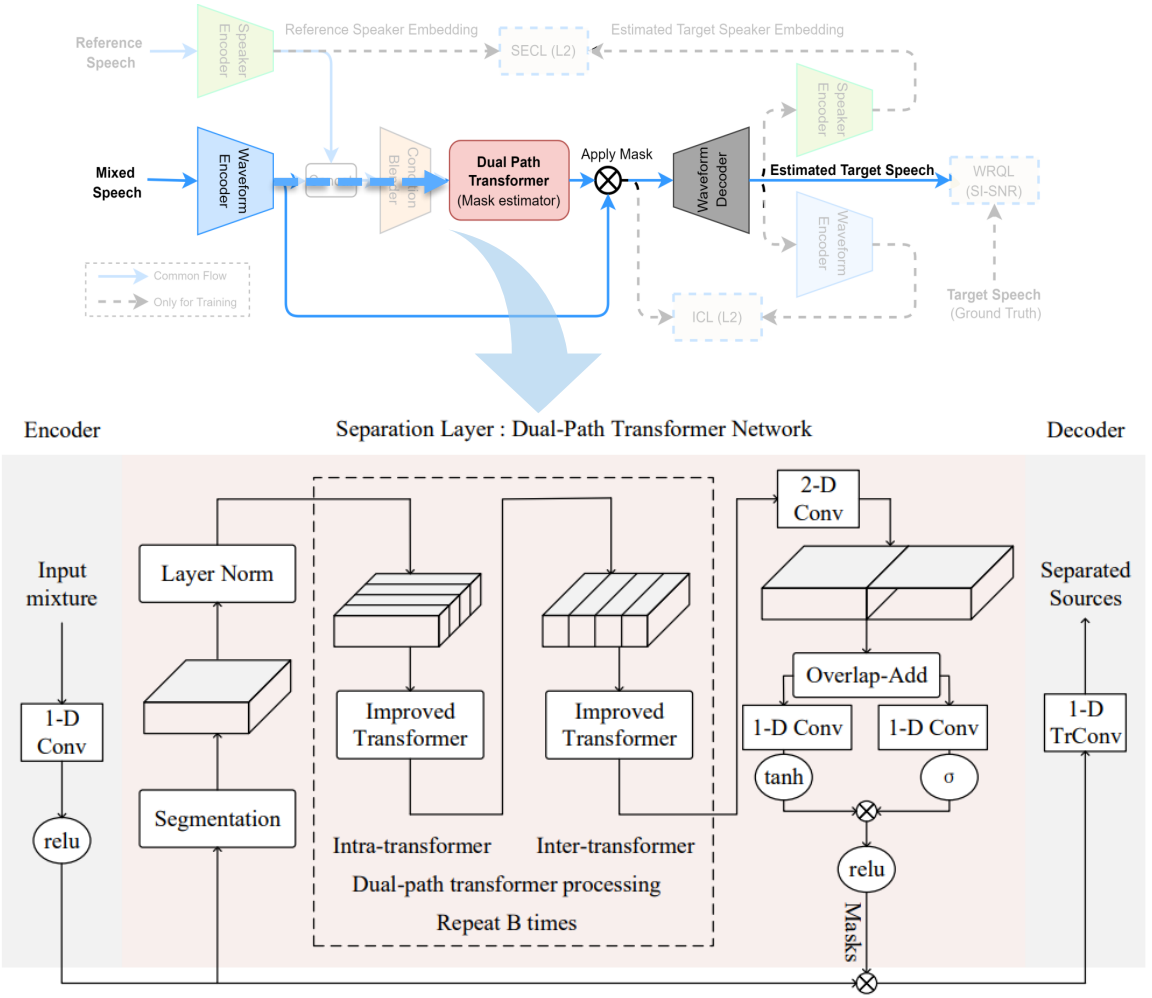


Method

Pipeline

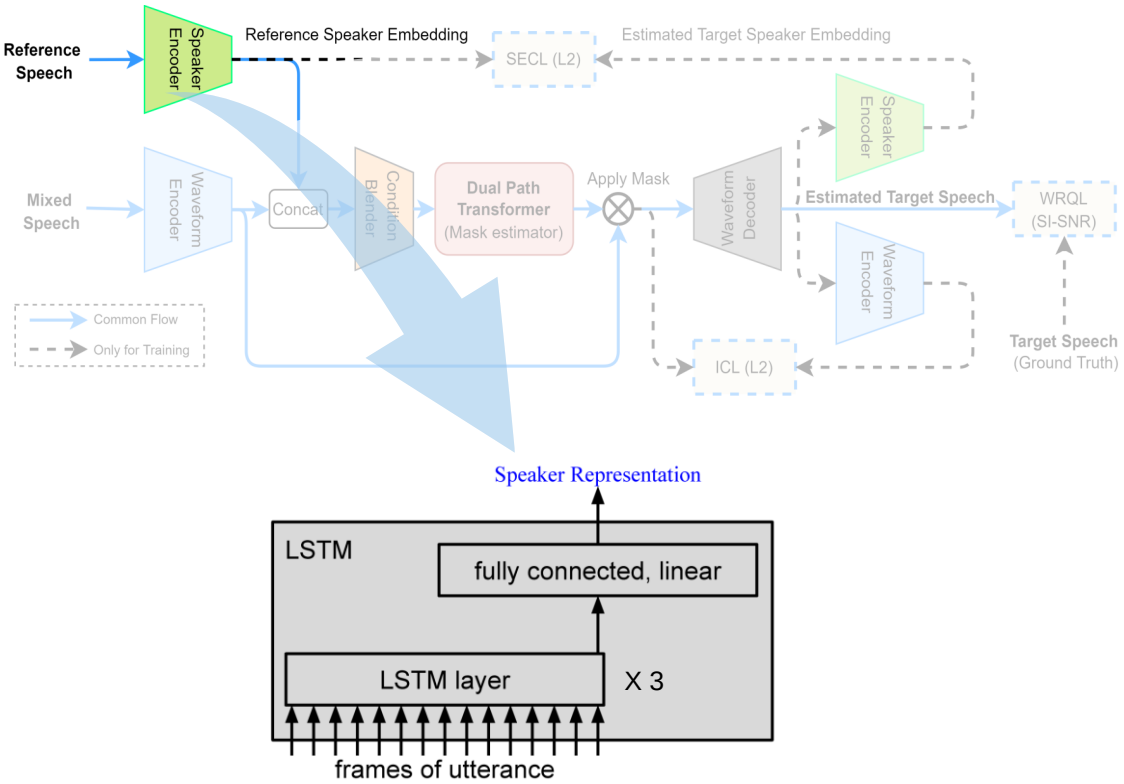

Reference Speech
Speaker Encoder
Reference Speaker Embedding
SECL (L2)
Estimated Target Speaker Embedding
Mixed Speech
Waveform Encoder
Concat
Condition Blender
Dual Path Transformer
(Mask estimator)
Apply Mask
Waveform Decoder
Estimated Target Speech
WRQL
(SI-SNR)
Speaker Encoder
Waveform Encoder
ICL (L2)
Target Speech
(Ground Truth)
Common Flow
Only for Training

Dual Path Transformer Backbone



[1]Dual Path Transformer Flowchart Source: Chen, Jingjing, Qirong Mao, and Dong Liu. "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation." *arXiv preprint arXiv*:2007.13975 (2020).

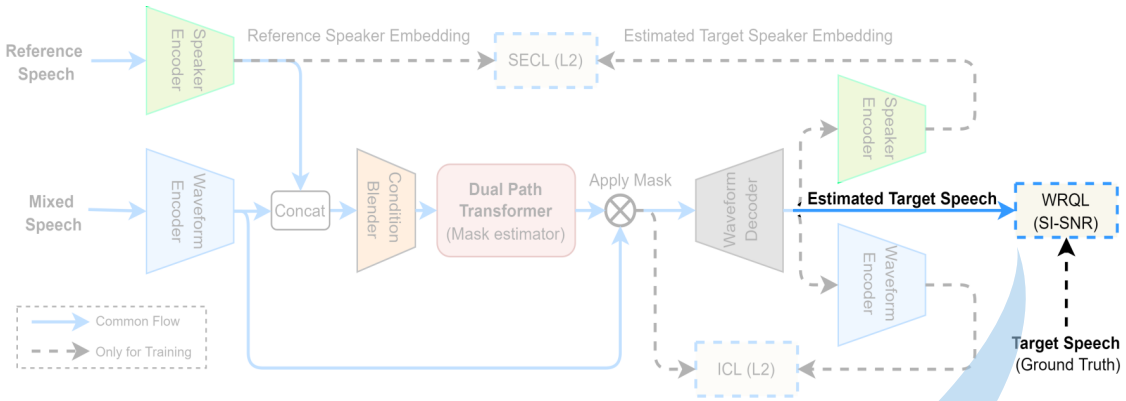
Speaker Encoder



[2]LSTM Image Source: Heigold, Georg, et al. "End-to-end text-dependent speaker verification." *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[3]Speaker Encoder architecture source: Wan, Li, et al. "Generalized end-to-end loss for speaker verification." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

Waveform Reconstruction Quality Loss (WRQL)



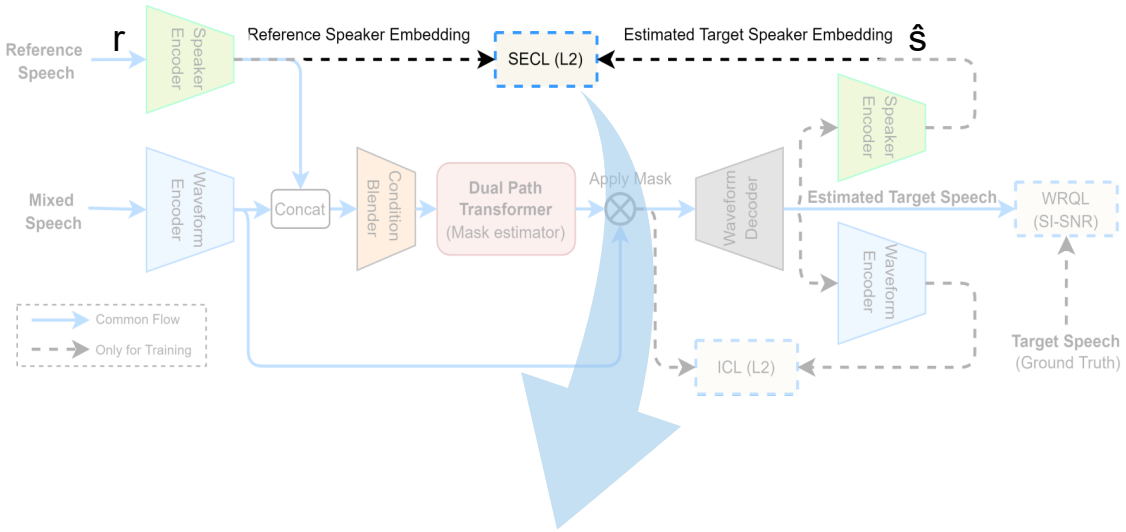
$$s_{\text{target}} = \frac{\langle \hat{s}, s \rangle s}{\|s\|^2}$$

$$e_{\text{noise}} = \hat{s} - s_{\text{target}}$$

$$SI\text{-}SNR := 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2}$$

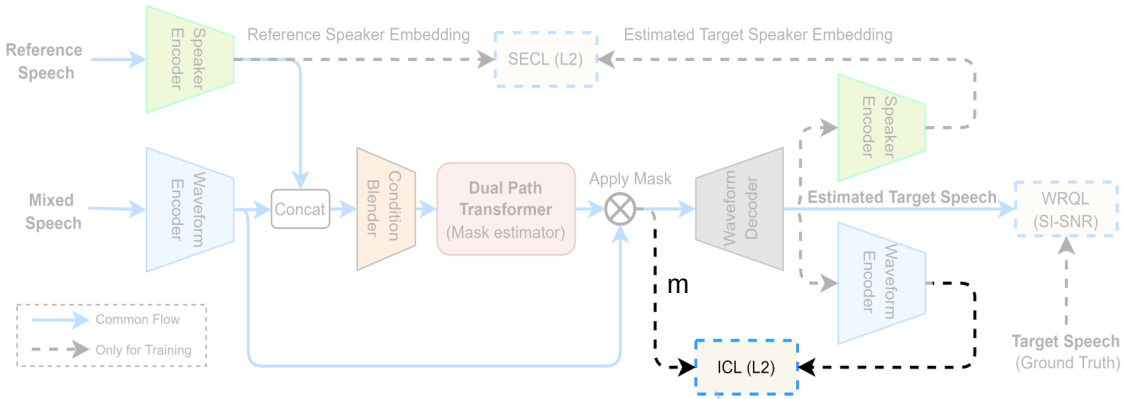
$$WRQL := -SI\text{-}SNR$$

Speker Embedding Consistancy Loss (SECL)



$$SECL := \|SE_\theta(r) - SE_\theta(\hat{s})\|^2$$

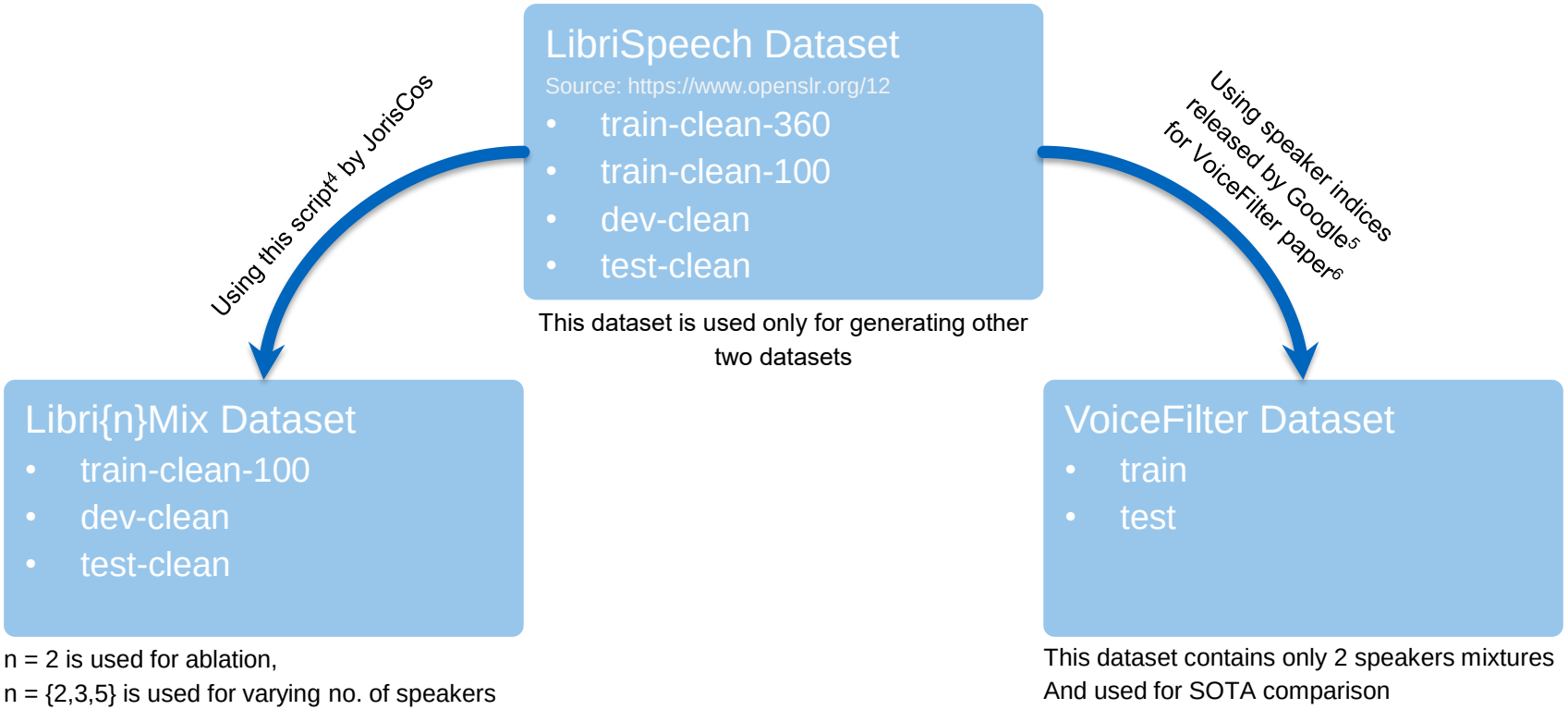
Inverse Consistancy Loss (ICL)



$$ICL := \|m - WE_\gamma(WD_\delta(m))\|^2$$

Results

Dataset



[4]Speaker https://github.com/JorisCos/LibriMix

[5]https://github.com/google/speaker-id/tree/master/publications/VoiceFilter/dataset/LibriSpeech

[6]Wang, Quan, et al. "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking." arXiv preprint arXiv:1810.04826 (2018).

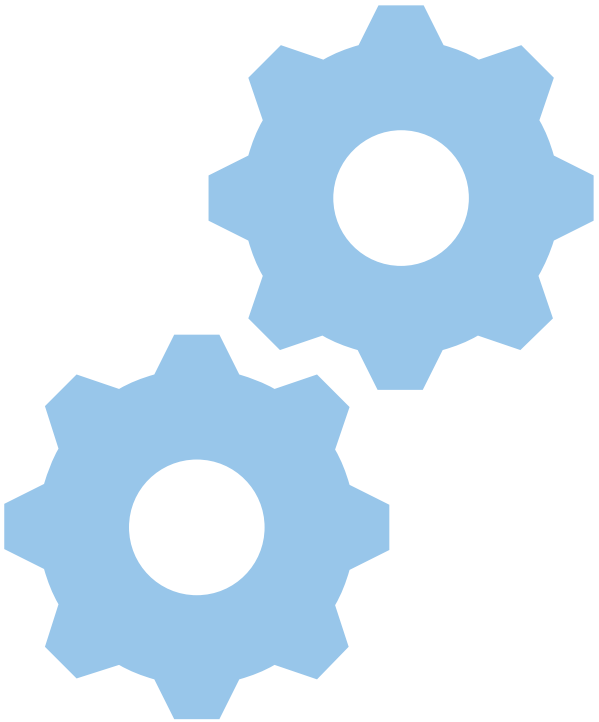
Experimental Setup

Environment:

- os: Ubuntu 16.04.7
- gpu: NVIDIA GTX1080Ti with CUDA 11.3
- python: 3.9.7
- pyTorch: 1.10.0

HyperParams:

- batch_size = 4
- learning_rate = 0.0001
- weight_decay = 1e-7
- no_of_attention_heads = 8
- optimizer: Adam
- reference_speech_sample_rate = 16 KHz
- mixed input and GT sample rate = 8 KHz
- audio_segment_length = 3 s



Gradual Development of the Model: Ablation

Model Variant	SDRi (dB)	SI-SNRi (dB)
Baseline*	11.13	10.42
Baseline+ICL	10.92	10.07
Baseline+ICL+SECL	10.95	10.15
Baseline+ICL+SECL+JointTraining	12.41	11.72
Spectron (with "DPTNet" backbone)	**13.94**	**13.23**

*__Baseline__ model refers to a system with fixed pretrained Speaker Encoder (from GE2E paper[7]) and "ConvTasnet"[8] backbone trained only with negative SI-SNR as WRQL

[7]Speaker Encoder architecture source: Wan, Li, et al. "Generalized end-to-end loss for speaker verification." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[8]Luo, Yi, and Nima Mesgarani. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation." IEEE/ACM transactions on audio, speech, and language processing 27.8 (2019): 1256-1266..

Spectron vs. State-of-the-Arts

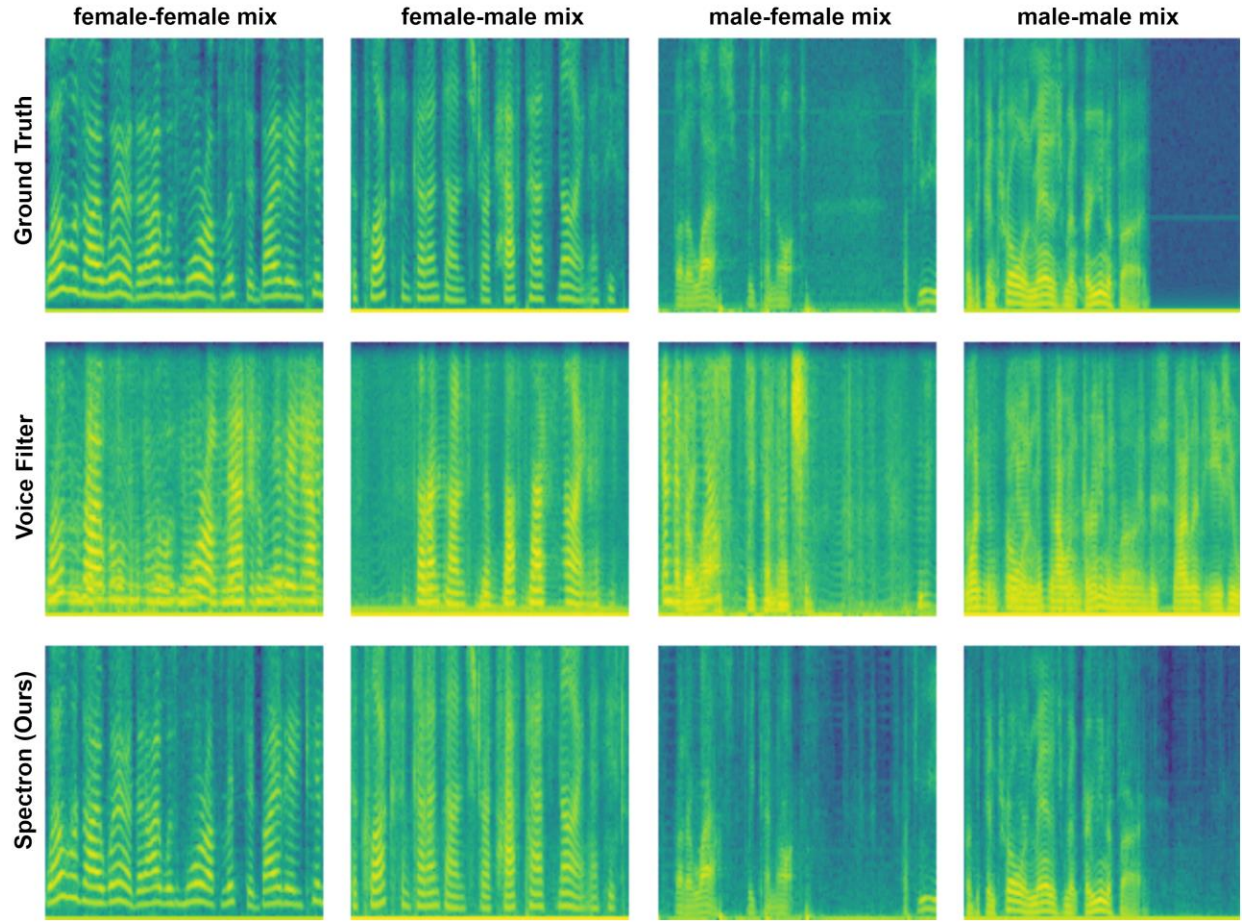
Model	SDRi (dB)	SI-SNRi (dB)
VoiceFilter[9]	7.8	-
AtssNet[10]	9.3	-
X-Tasnet[11]	13.8	12.7
X-Tasnet with LoD[11]	14.7	13.8
Spectron (ours)	13.9	12.8

[9]Wang, Quan, et al. "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking." arXiv preprint arXiv:1810.04826 (2018).

[10]Li, Tingle, et al. "Atss-net: Target speaker separation via attention-based neural network." *arXiv preprint arXiv:2005.09200* (2020).

[11]Zhang, Zining, Bingsheng He, and Zhenjie Zhang. "X-tasnet: Robust and accurate time-domain speaker extraction network." *arXiv preprint arXiv:2010.12766* (2020).

Spectron vs. Voice Filter*: Qualitative Comparison



*Wang, Quan, et al. "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking." arXiv preprint arXiv:1810.04826 (2018).

Varrying Number of Speech Sources

Separately Trained

#Speakers	SDRi (dB)	SI-SNRi (dB)
2	13.94	13.23
3	10.76	9.89
5	5.62	4.07

Trained together in Mixed Batches

#Speakers	SDRi (dB)	SI-SNRi (dB)
2	13.45	12.60
3	11.40	10.29
5	8.25	6.71

Live Demo!

Conclusion

Conclusion and Future Scope

In summary:

- A Transformer based Speaker extraction Framework
- Two additional novel objective functions
- Joint training strategy, along with above two points, improves baseline
- For n-mix case, mixed batch training is better than separate training

In future:

- Transformer based Speaker Encoder
- Speaker Presence Invariant Training
- Down stream applications

Thank you!

Any questions?