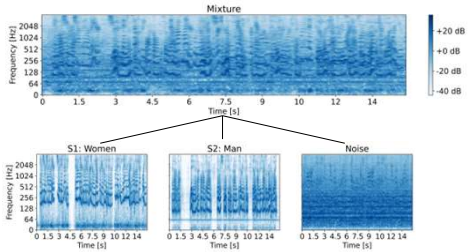
Denoising in Dual-Path-RNN-based Speech Separation

Anna Grillenberger, Daniel Schober, Pablo Taboas Rivas, Huiyu Lu
02456 Deep Learning

Introduction

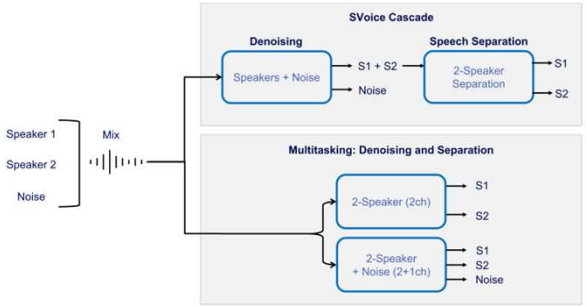
Why is speech separation relevant?
Many situations in life require the ability of separating a single voice from a conversation of multiple people:
- Speech assistants, voice messaging, group calls
- Hearing aids: Separating a single conversation in a crowded and noisy environments ("cocktail-party-problem")



Our goal – Research question

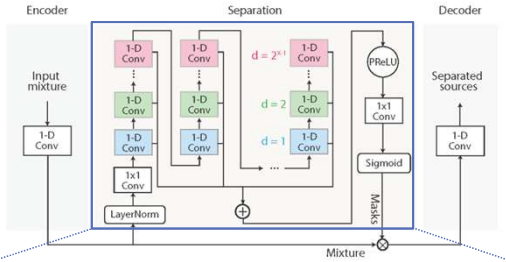
Challenges of Speech Separation
- Unknown number of speakers
- Background noise
- Long-sequence and Time-domain

Since the SVoice paper focuses on the problem of separating an unknown number of multiple speakers, we will address the second problem of speech separation – **how to deal with noise**.
Taking the SVoice model architecture as a base, we will research the following:
- How do different architectures affect the quality of speech separation in noisy environments?
  - Performing denoising and speech separation simultaneously
    - Only speakers as output channels (SVoice)
    - Noise as an additional output channel
  - Splitting the tasks – First denoising and then speech separation
- How does the model perform on different types of noise?
  - Different types of background noise
  - Different intensities of noise

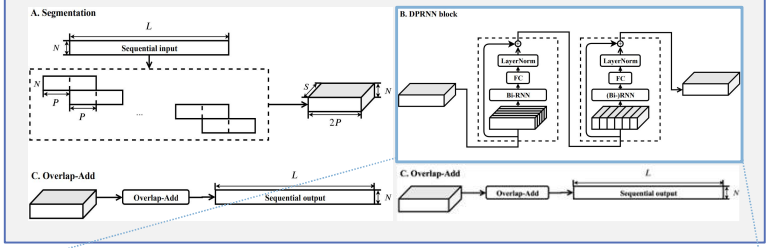


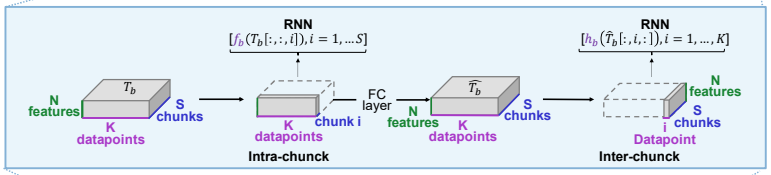
Architecture

TasNET (starting point)

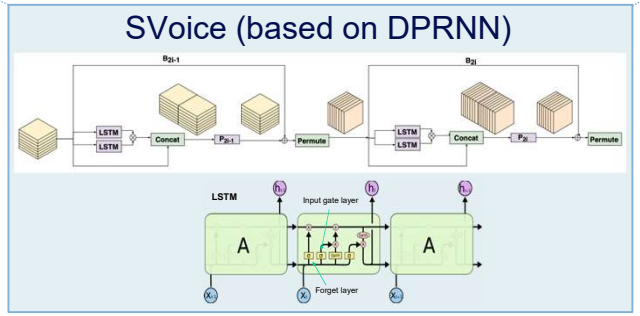


DPRNN (based on TasNET)



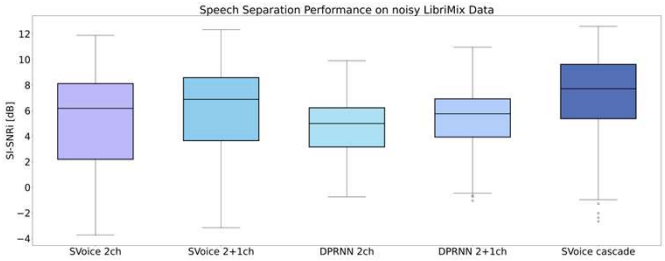


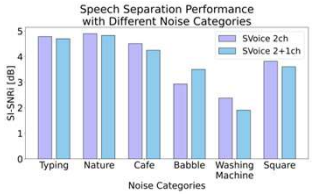
SVoice (based on DPRNN)

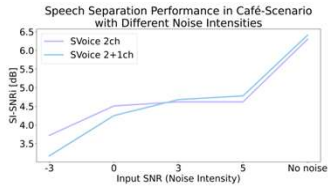


Results

Training Information
- Dataset Model Comparison: LibriMix (2 speakers with ambient noise samples)
  - 800 training, 200 validation samples
- Dataset Noise Testing: 2 speakers of LibriMix + categorical noise
  - 20 samples per category, MS-SNSD
- Sample rate: 8000 Hz, duration: 4-6 seconds
- Mono input, waveform domain
- Number of model parameters
  - DPRNN: 3.6 million
  - SVoice: 7.5 million







Discussion

- For known noise, models trained with an additional noise channel outperform models without
- For unknown noise, the results of noise-channel-models depend heavily on the noise category
  → It performs good for the noise category it was trained on (background speakers)
- Svoice outperforms DPRNN on a new dataset (LibriMix)
- The 2+1 channel model usually outperforms the 2 channel model for low noise intensities
- Multitask-learning has been proven to be possible. However, the cascade approach outperforms it slightly in exchange with more computation

Future work: Increase the variety of noises for training and the SNR between the signals. Evaluate the cascade approach in one model.

References

[1] Luo, Yi, Zhuo Chen, and Takuya Yoshioka. "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation." ICASSP (2019).
[2] Eliya Nachmani, Yossi Adi, and Lior Wolf. 2020. "Voice separation with an unknown number of multiple speakers." In Proceedings of the 37th International Conference on Machine Learning (ICML'20).
[3] 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.
[4] Yuan-Kuei Wu, Chao-I Tuan, Hung-yi Lee, and Yu Tsao. "SADDEL: Joint Speech Separation and Denoising Model based on Multitask Learning." CoRR (2020)