Current Trends in Audio Source Separation

Fabian-Robert Stöter
INRIA

Stefan Uhlich
Sony R&D Center

AES Virtual Symposium: Applications of Machine Learning in Audio

September 28th, 2020

Contents

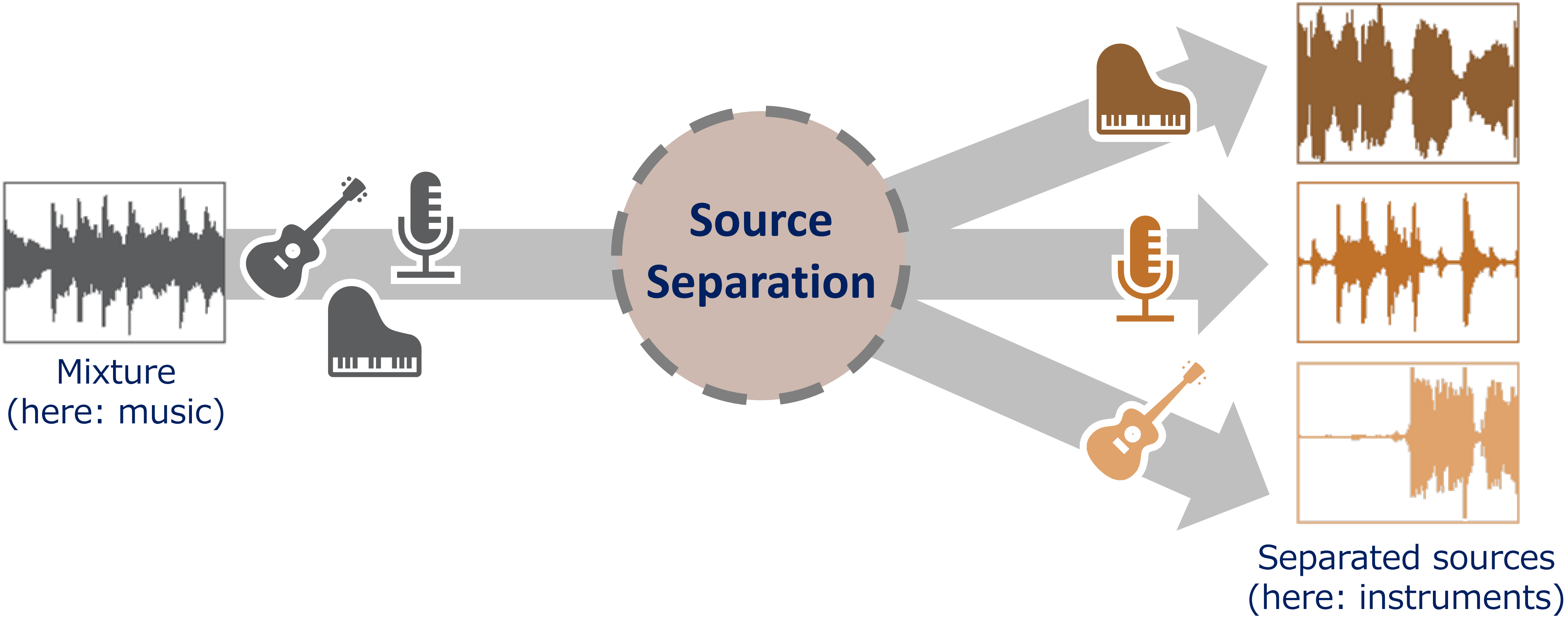
- Introduction to Audio Source Separation
- Current Trends and Open Problems
- Ecosystems, Datasets and Upcoming Competitions

Contents

- Introduction to Audio Source Separation
- Current Trends and Open Problems
- Ecosystems, Datasets and Upcoming Competitions

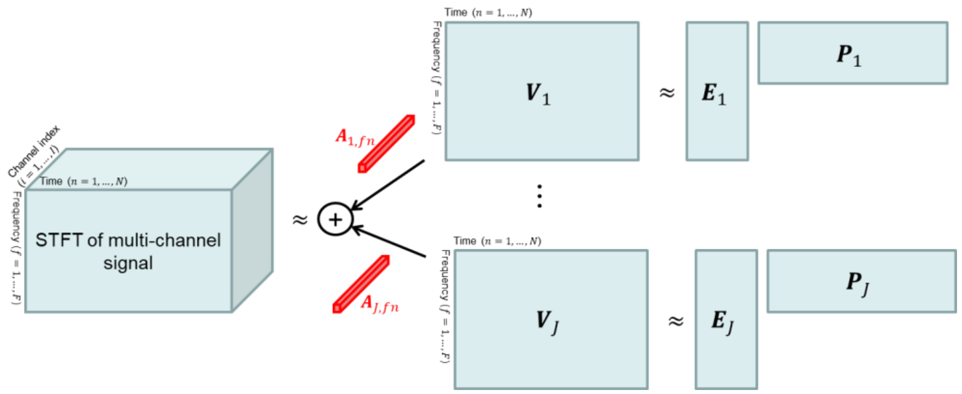
Introduction to Audio Source Separation (I)

Task: Separate mixture into individual sound sources



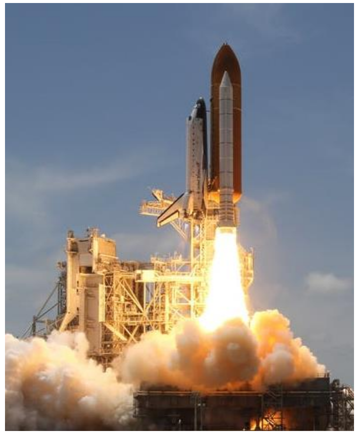
Introduction to Audio Source Separation (II)

- In general, this task is ill-posed and difficult to solve
  - We often deal with an under-determined source separation problem
    - E.g., monaural speech enhancement (single sample $x(n)$ for two sources)
    - E.g., stereo music separation (single sample $\boldsymbol{x}(n)$ for four sources)
  - Classical methods only worked to some extent
    - Best method for music by 2012 was multichannel NMF (FASST), see e.g. [1]



[1] Ozerov, Alexey, et al. "A general flexible framework for the handling of prior information in audio source separation." TASLP 2011.

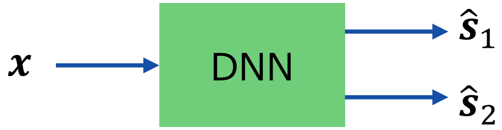
Introduction to Audio Source Separation (III)



- However, DNNs came to a rescue ☺
  - First approaches: [1-4]
  - Popular architectures nowadays
    - Time domain: Conv-TasNet, DPRNN, Demucs, ⋯
    - Frequency domain: Open-Unmix, Deep U-Net, ⋯

    - Same architecture can be used for separation of different/same source types
      - Loss function for different source types (e.g., vocals/accompaniment)

$$L(\hat{\boldsymbol{s}}_1, \boldsymbol{s}_1) + L(\hat{\boldsymbol{s}}_2, \boldsymbol{s}_2)$$

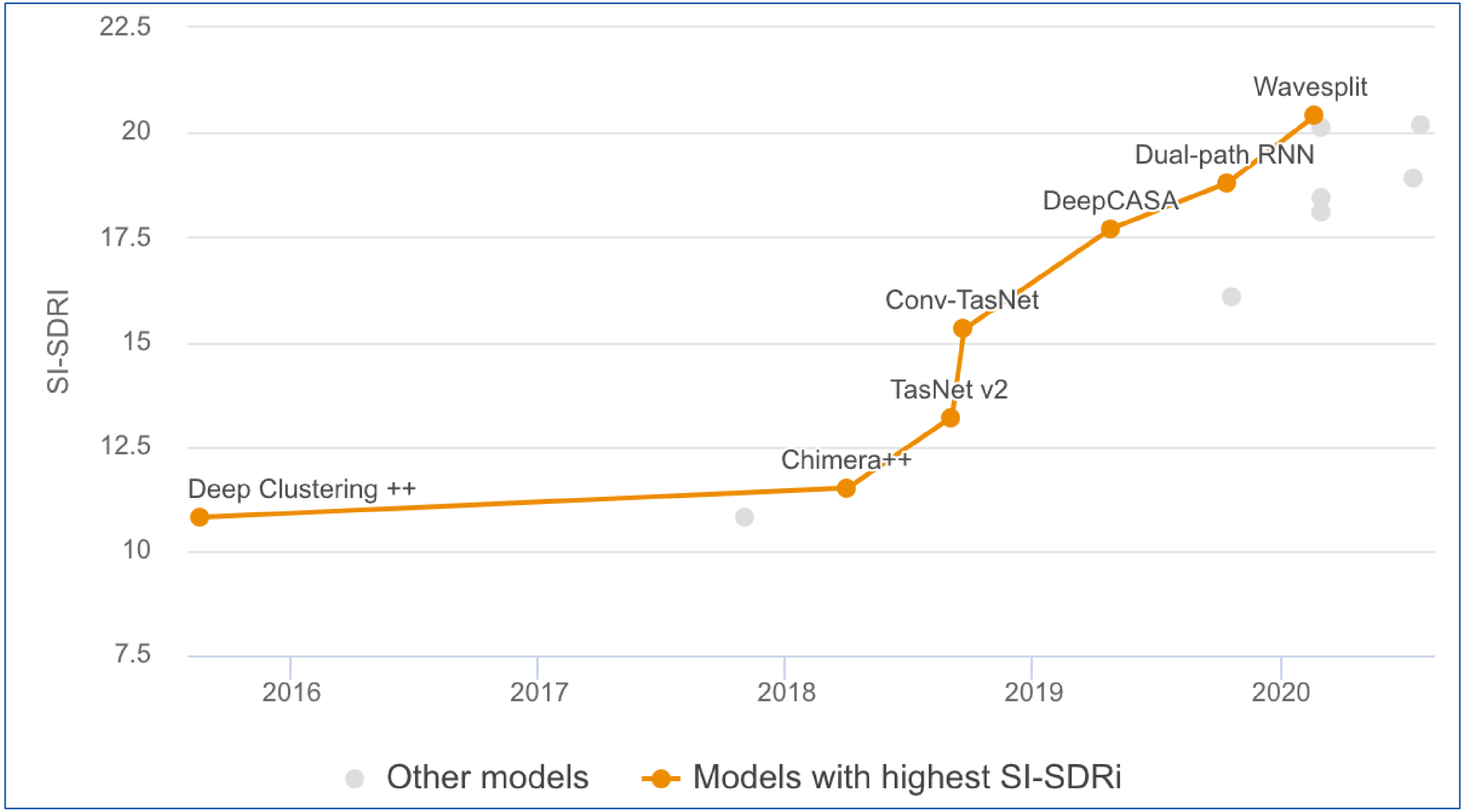


      - PIT loss function [5] for same source types (e.g., two unknown speakers or two violins)

$$\min\{L(\hat{\boldsymbol{s}}_1, \boldsymbol{s}_1) + L(\hat{\boldsymbol{s}}_2, \boldsymbol{s}_2), L(\hat{\boldsymbol{s}}_1, \boldsymbol{s}_2) + L(\hat{\boldsymbol{s}}_2, \boldsymbol{s}_1)\}$$

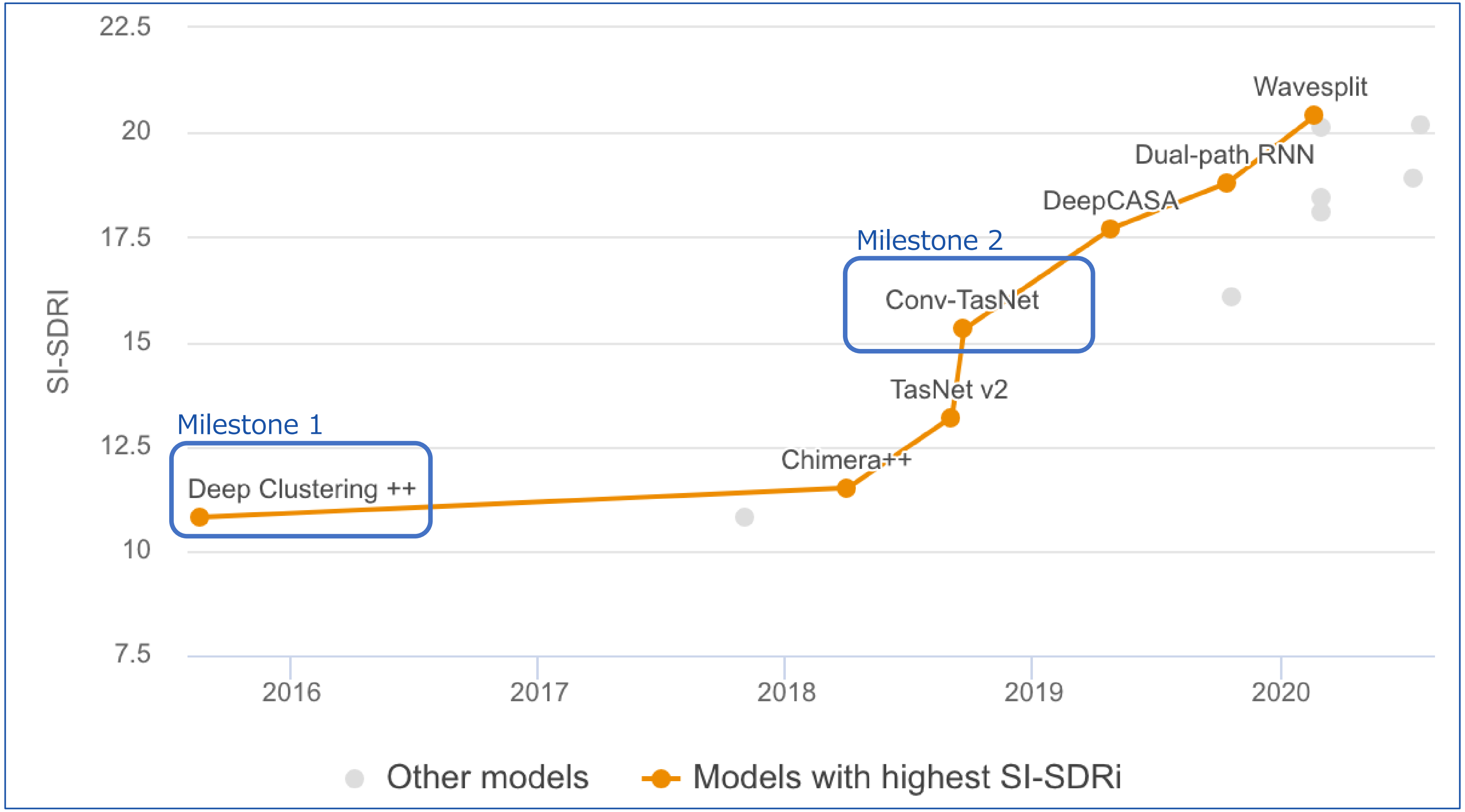
[1] Narayanan, Arun, and DeLiang Wang. "Ideal ratio mask estimation using deep neural networks for robust speech recognition." ICASSP 2013.
[2] Huang, Po-Sen, et al. "Deep learning for monaural speech separation." ICASSP 2014.
[3] Grais, Emad M., et al. "Deep neural networks for single channel source separation." ICASSP 2014
[4] Weninger, Felix, et al. "Discriminatively trained recurrent neural networks for single-channel speech separation." GlobalSIP 2014.
[5] Yu, Dong, et al. "Permutation invariant training of deep models for speaker-independent multi-talker speech separation." ICASSP 2017.

Progress for Speech Separation



Plot shows scale-invariant SDR improvement (SI-SDRi) on wsj0-mix2 (higher is better)
Source: https://paperswithcode.com/task/speech-separation

Progress for Speech Separation

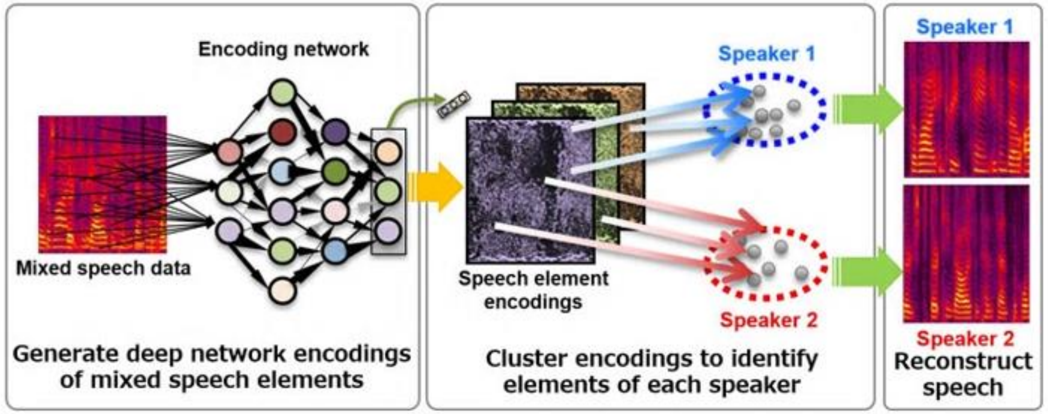


Plot shows scale-invariant SDR improvement (SI-SDRi) on wsj0-mix2 (higher is better)
Source: https://paperswithcode.com/task/speech-separation

Two milestones

- Milestone 1: DL-based approach for speech separation
  - Deep Clustering [1]
  - Permutation invariant training (PIT) [2]

$$\min\{L(\hat{\boldsymbol{s}}_1,\boldsymbol{s}_1)+L(\hat{\boldsymbol{s}}_2,\boldsymbol{s}_2), L(\hat{\boldsymbol{s}}_1,\boldsymbol{s}_2)+L(\hat{\boldsymbol{s}}_2,\boldsymbol{s}_1)\}$$



Source: https://de.mitsubishielectric.com/en/news-events/releases/global/2017/0524-e/index.html

- Milestone 2: End-to-end time domain architectures
  - TasNet [3] / Conv-TasNet [4]

Demo: Comparison of Deep Clustering and Conv-TasNet

[1] Hershey, John R., et al. "Deep clustering: Discriminative embeddings for segmentation and separation." ICASSP 2016
[2] Yu, Dong, et al. "Permutation invariant training of deep models for speaker-independent multi-talker speech separation." ICASSP 2017
[3] Luo, Yi, and Nima Mesgarani. "Tasnet: time-domain audio separation network for real-time, single-channel speech separation." ICASSP 2018.
[4] Luo, Yi, and Nima Mesgarani. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation." TASLP 2019

General Source Separation architecture

- Encoder
  - Fixed filterbanks (STFT,MEL)
  - Trainable filterbanks
  - Free filterbanks
- Separator
  - LSTMs, TCN, U-Net...
  - Complex networks (CaC)
- Decoder
  - Inverse encoders
  - Vocoder (waveglow?)
- Loss
  - fixed/ permutation invariant
  - Spectrogram loss, end2end loss, combined

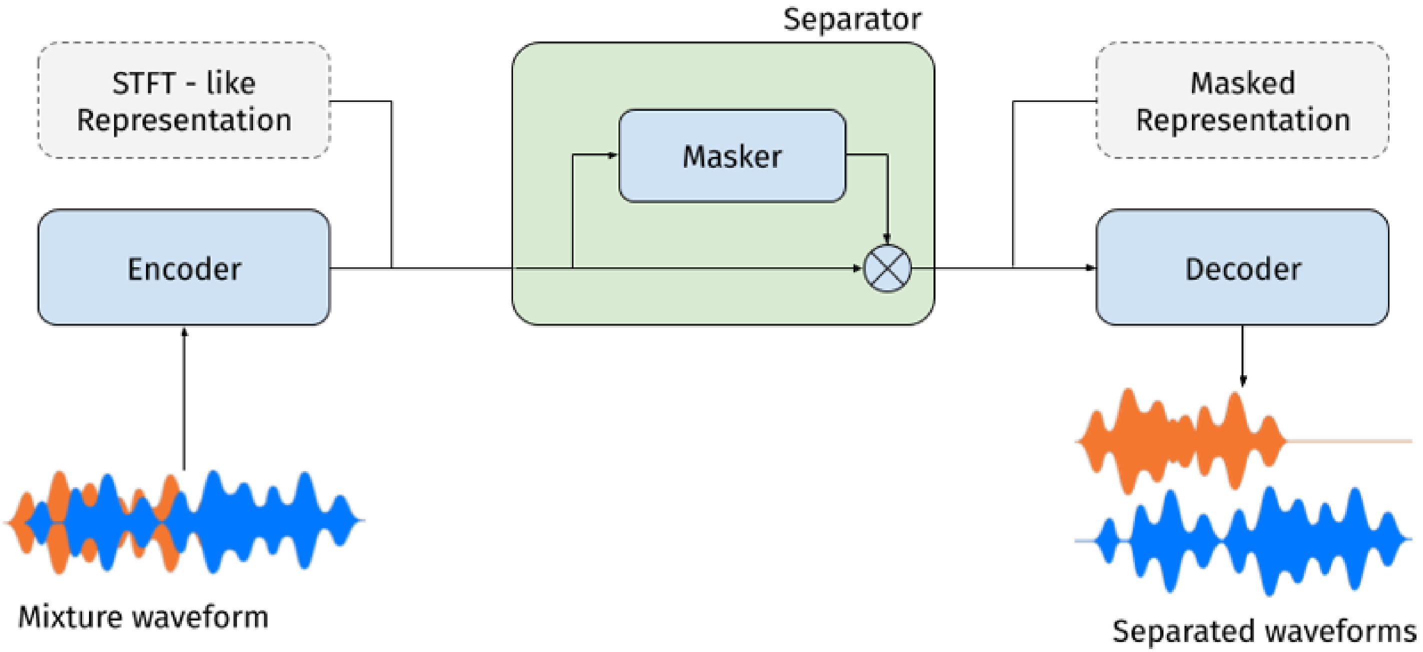
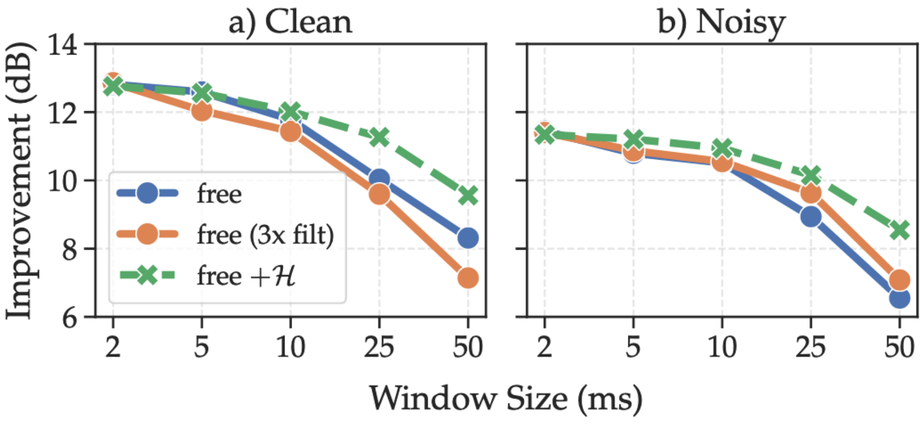


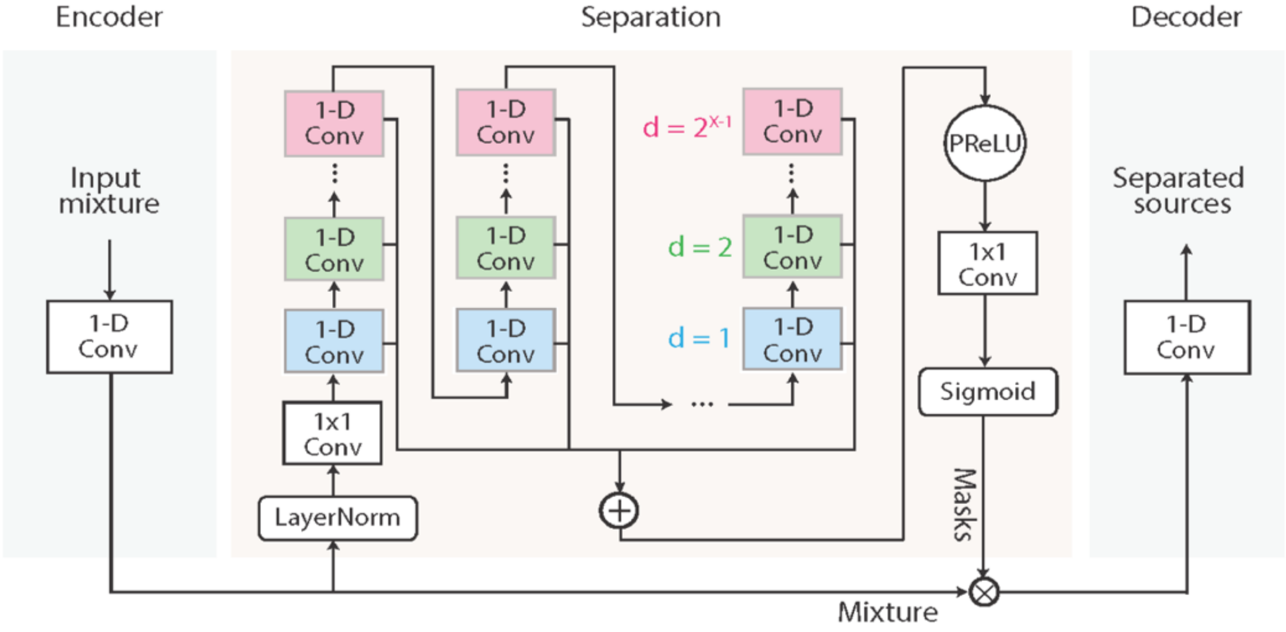
Figure from Pariente, Manuel, et al. "Asteroid: the PyTorch-based audio source separation toolkit for researchers." *arXiv preprint arXiv:2005.04132* (2020).

Reference Architecture: Conv-TasNet

- All improvements add up 5dB
  - Learnable transforms
  - Time domain loss
  - Normalization (layer norm)
  - Separator (TCN capacity)
  - Short windows (almost 2dB)



M Pariente, S Cornell, A Deleforge, E Vincent "Filterbank design for end-to-end speech separation"- ICASSP 2020



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.

Contents

- Introduction to Audio Source Separation
- Current Trends and Open Problems
- Ecosystems, Datasets and Upcoming Competitions

Trend #1: From “Supervised” to “Universal” Separation

Supervised separation

- Mixture is composed of a small number of sources (2 or 4)
- Number of sources known a priori
- Type of sources known a priori

…

Universal separation

- Unknown number of sources
- Unknown type of sources

Example datasets
- MSD100/DSD100/MUSDB18
- wsj0-mix2

Example datasets
- WildMix dataset [1]
- Free Universal Sound Separation (FUSS) dataset [2]
- Slakh2100 [3]

[1] Zadeh, Amir, et al. "WildMix Dataset and Spectro-Temporal Transformer Model for Monoaural Audio Source Separation." *arXiv preprint arXiv:1911.09783* (2019).
[2] https://opensource.googleblog.com/2020/04/free-universal-sound-separation.html
[3] Manilow, Ethan, et al. "Cutting music source separation some slakh: a dataset to study the impact of training data quality and quantity." WASPAA 2019.

Trend #1: From "Supervised" to "Universal" Separation

- Universal sound separation not yet solved but recently lot of progress
  - Source separation for an unknown number of sources [1, 2]
    - Example: Approach [1] uses "one vs. rest"-PIT $L = \min_i l(\hat{s}(t), s_i(t)) + \frac{1}{N-1} l(\hat{r}(t), \sum_{n \neq i} s_n(t))$

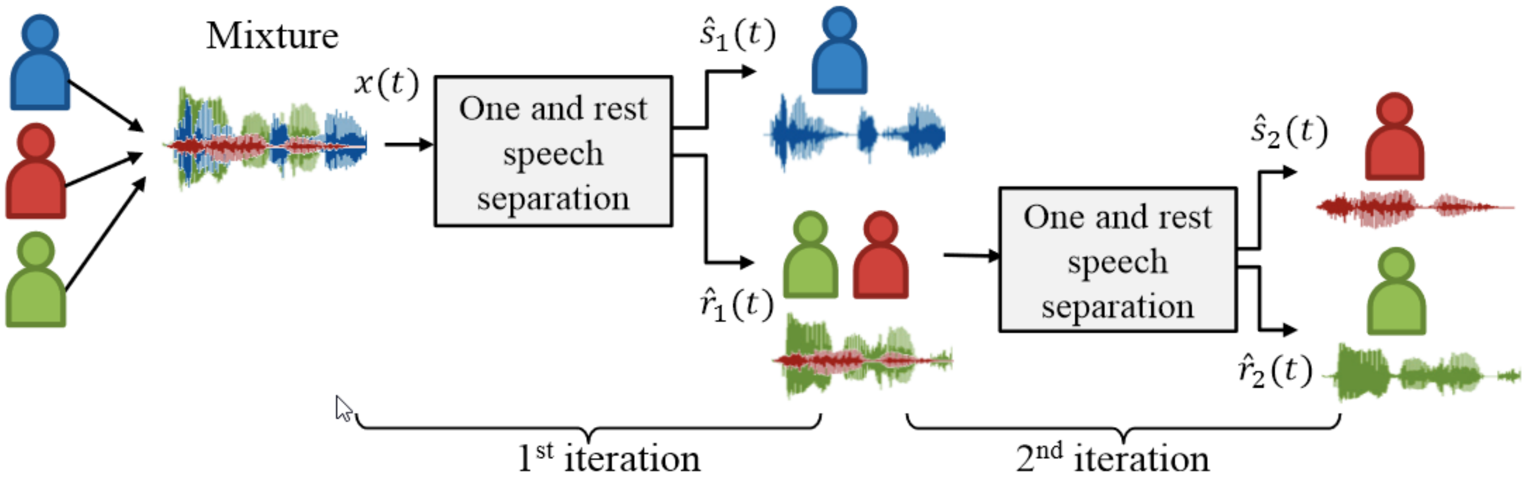


Figure from [1]

[1] Takahashi, Naoya, et al. "Recursive speech separation for unknown number of speakers." InterSpeech 2019
[2] Nachmani, Eliya, Yossi Adi, and Lior Wolf. "Voice Separation with an Unknown Number of Multiple Speakers." ICML 2020.

Trend #1: From "Supervised" to "Universal" Separation

- Universal sound separation not yet solved but recently lot of progress
  - Source separation for large number of classes [1–3]
    - Example: System [1] for the 527 AudioSet classes
      - Can separate mixtures of two sources using a conditioned network $f(\boldsymbol{x}, \boldsymbol{c})$
      - Training with weakly-labeled data is done in two steps:
        1. Train sound event detector (SED) on AudioSet
        2. Use SED to detect anchor segments which can be used for training

$$f(\boldsymbol{s}_1 + \boldsymbol{s}_2, \boldsymbol{c}_j) = \hat{\boldsymbol{s}}_j$$

where condition vector contains probabilities from SED

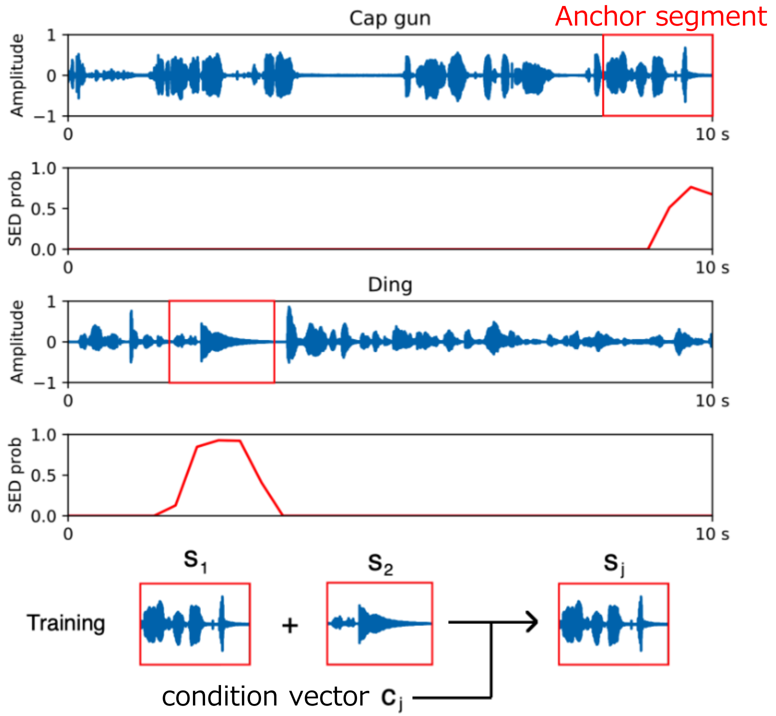


Figure from [1]

  - Training on dataset with only mixtures: Mixture invariant training (MixIT) [3]

[1] Kong, Qiuqiang, et al. "Source separation with weakly labelled data: An approach to computational auditory scene analysis." ICASSP 2020.
[2] Kavalerov, Ilya, et al. "Universal sound separation." WASPAA 2019.
[3] Wisdom, Scott, et al. "Unsupervised sound separation using mixtures of mixtures." *arXiv preprint arXiv:2006.12701* (2020).

Trend #2: Dealing with “Imperfect” Training Data

- “Imperfect” = weakly labeled data
  - Idea of [1, 2] is to use sound event classifier
    - Classifier provides labels on which source is active (frame-level or clip-level)
    - This information is used to train the separation network
  - Approach [1]:

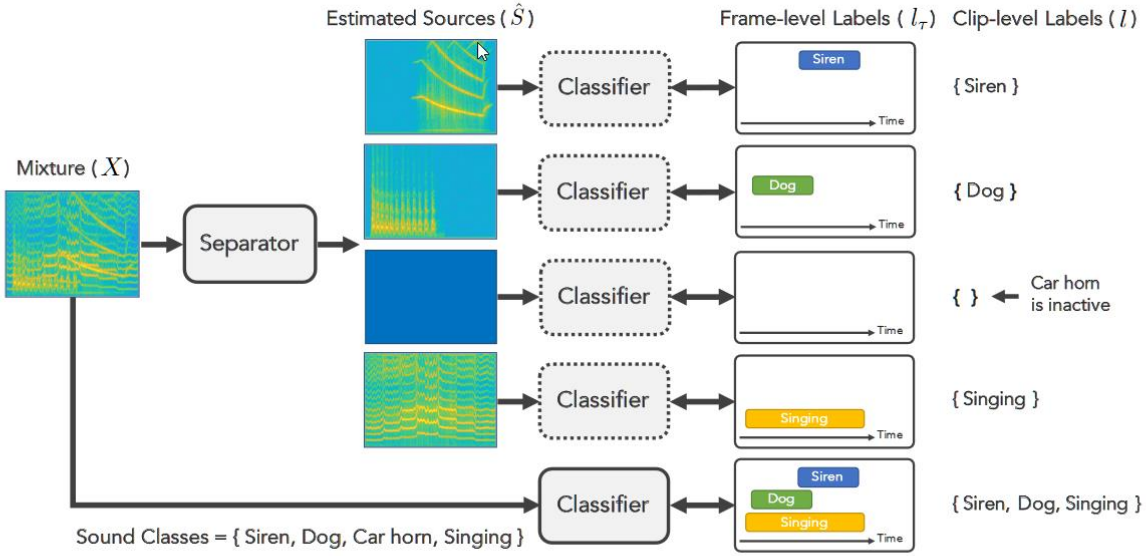
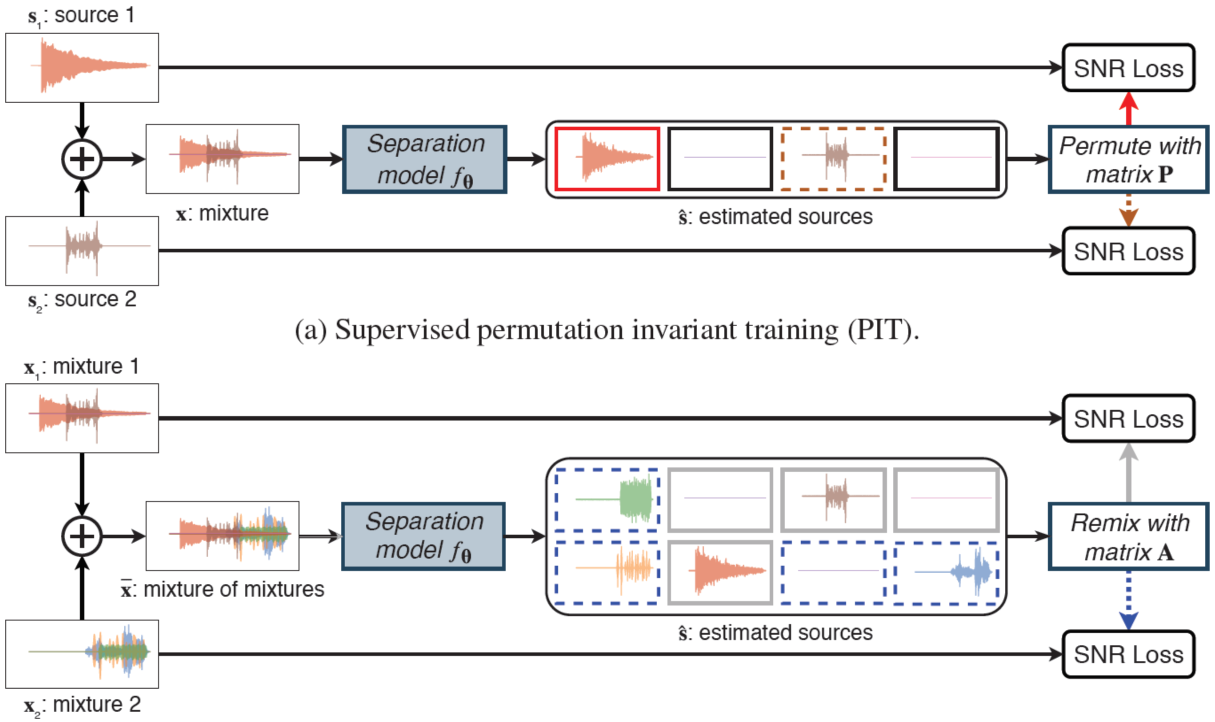


Figure from [1]

[1] Pishdadian, Fatemeh et al. "Finding strength in weakness: Learning to separate sounds with weak supervision." TASLP 2020.
[2] Kong, Qiuqiang, et al. "Source separation with weakly labelled data: An approach to computational auditory scene analysis." ICASSP 2020.

Trend #2: Dealing with “Imperfect” Training Data

- “Imperfect” = unknown mixtures of sources
  - Separation net can be learned even in this unsupervised setting using MixIT [1] (assuming that mixtures are “random”)



(a) Supervised permutation invariant training (PIT).

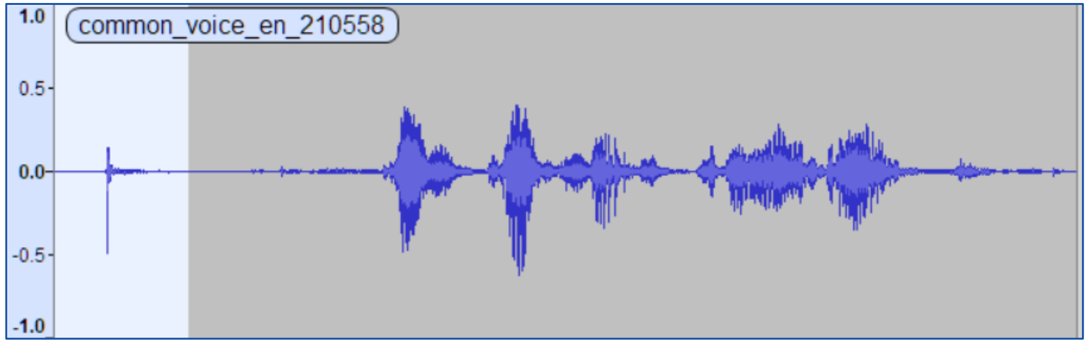
(b) Unsupervised mixture invariant training (MixIT).

Figure from [1]

[1] Wisdom, Scott, et al. "Unsupervised sound separation using mixtures of mixtures." *arXiv preprint arXiv:2006.12701* (2020).

Trend #2: Dealing with "Imperfect" Training Data

- "Imperfect" = noisy speech
  - For speech enhancement, obtaining large amounts of high-quality, multi-language clean speech is not straight-forward
  - Many speech datasets were created for ASR
    - Might still contain noise (maybe even on purpose)
    - E.g., Mozilla Common Voice:
      - Microphone switch-on sound
      - sample dropping
      - …



  - Some work on this topic using "Noise2Noise" approach [1, 2] or bootstrapping [3]
  - But problem is still unsolved – MixIT cannot be used as "mixture" is not random

[1] Chang, Yen-Yu et al. "Noise-to-noise speech enhancement: speech denoising without clean speech.", 2019
[2] Alamdari, Nasim et al. "Improving Deep Speech Denoising by Noisy2Noisy Signal Mapping." arXiv preprint arXiv:1904.12069 (2019).
[3] Wang, Yu-Che et al. "Self-supervised Learning for Speech Enhancement." ICML 2020.

Trend #3: Perceptual Loss Functions for SE

- Perceptual measures are commonly used for speech enhancement
  - E.g., PESQ, composite measures (CSIG, CBAK, COVL) and STOI
  - Recent work has taken these to define perceptual loss functions, e.g. [1-5]
    - Either approximating them by differentiable terms or by treating them as black-box (finite differences, QualityNet, MetricGAN, ⋯)
    - Helps to improve perceptual quality of the separated speech

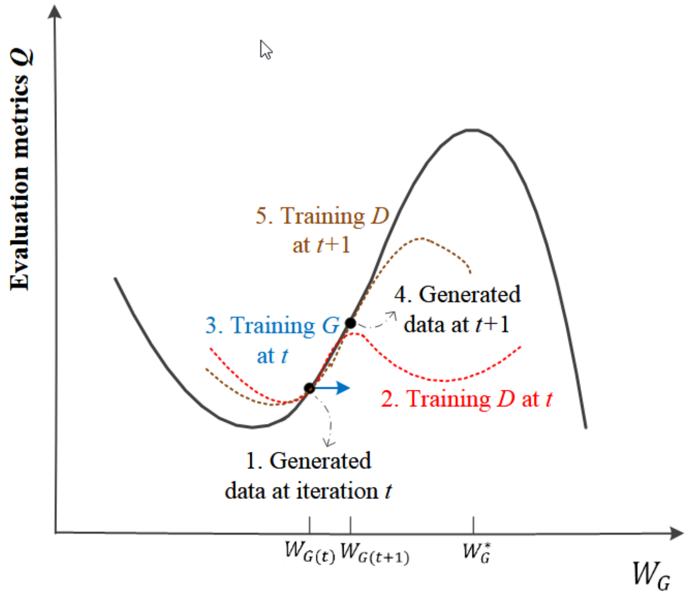


Figure from [4]

[1] Martín-Doñas, et al. "A deep learning loss function based on the perceptual evaluation of the speech quality." SPL 2018
[2] Zhang, Hui, et al. "Training supervised speech separation system to improve STOI and PESQ directly." ICASSP 2018.
[3] Fu, Szu-Wei, et al. "Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality." SPL 2019
[4] Fu, Szu-Wei, et al. "MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement." ICML 2019
[5] Kolbæk, Morten, et al. "On loss functions for supervised monaural time-domain speech enhancement." TASLP 2020

Trend #3: Perceptual Loss Functions for SE

- However for music, perceptual measures are not really common
  - PEASS seldomly used although APS showed good correlation with humans in [1]
  - Very recently an interesting idea came up in [2] for speech/vocals separation
    - A codec $\Phi(x)$ (e.g., low-bit rate MP3) is used to define a better loss function

$L(s,\hat{s})$ → Perceptual loss → $L\big(\Phi(s),\Phi(\hat{s})\big)$ → Differentiability → $L\big(g_\Phi(s),g_\Phi(\hat{s})\big)$

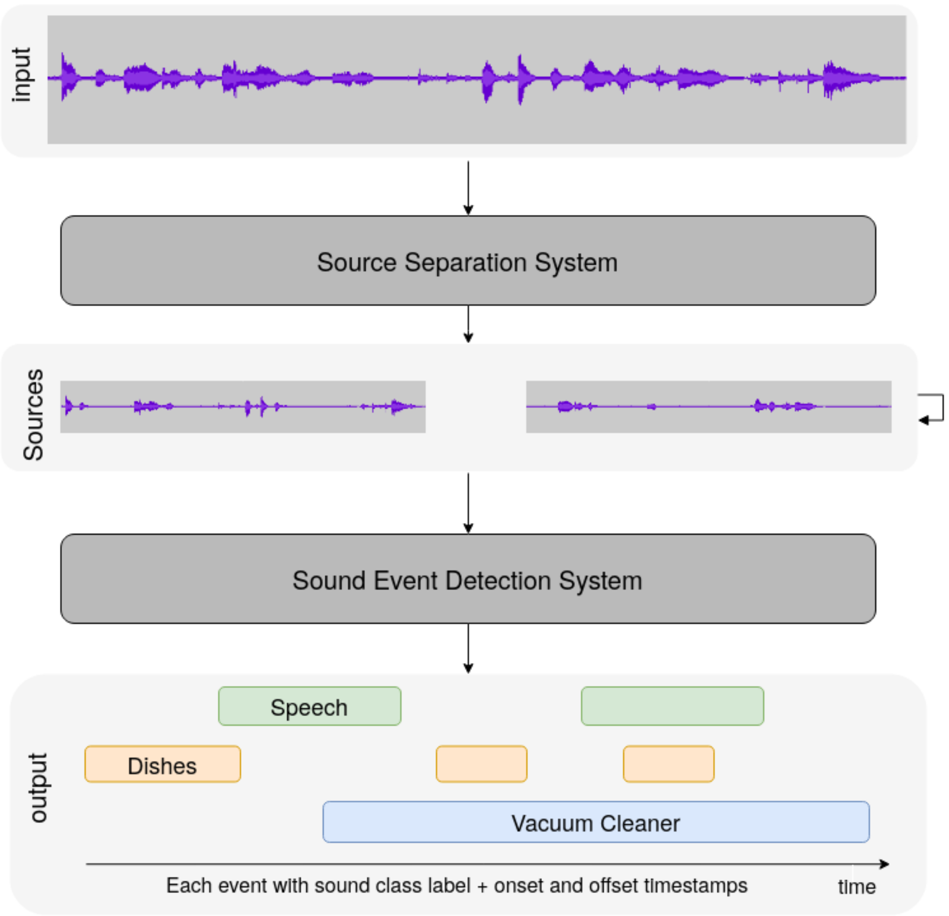
  - There is a need for more work on perceptual measures/loss functions for music

Demo: Comparison of artefacts for Conv-TasNet and Open-Unmix

[1] Ward, Dominic, et al. "BSS Eval or PEASS? Predicting the perception of singing-voice separation." ICASSP 2018.
[2] Ananthabhotla, Ishwarya et al. "Using a Neural Network Codec Approximation Loss to Improve Source Separation Performance in Limited Capacity Networks." WCCI/IJCNN 2020

Trend #4: Multi-Task Training

- Separation + Classification
  - Does separation help classification task?
    - Long standing research question
    - Relevant for music and speech tasks
  - Results will be presented at DCASE 2020



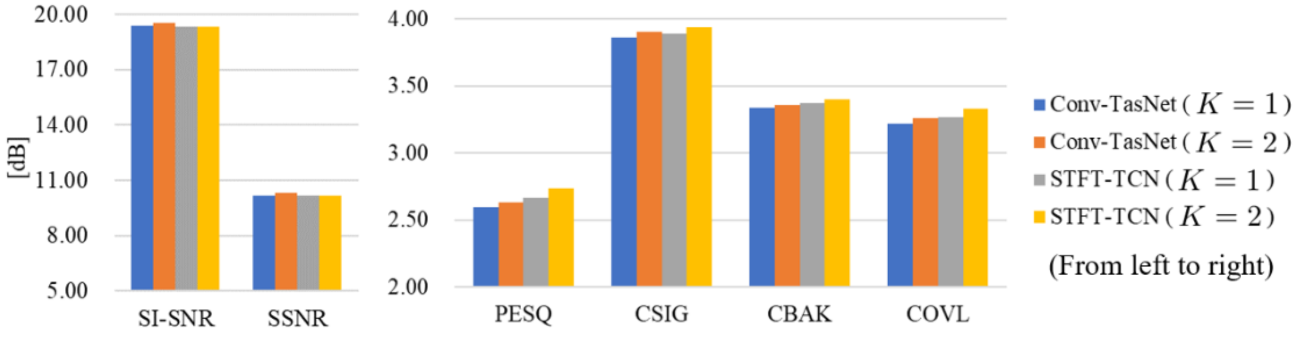
http://dcase.community/challenge2020/task-sound-event-detection-and-separation-in-domestic-environments

Trend #4: Multi-Task Training

- Multi-task training for speech enhancement
  - Training jointly mask for speech and noise is better than only speech mask [1, 2]

$$L = L(\boldsymbol{x}, \hat{\boldsymbol{x}}) + L(\boldsymbol{n}, \hat{\boldsymbol{n}})$$

  - Additional task acts as "regularizer"
  - Speech mask generalizes better



Results from [2] for training only speech mask ($K = 1$) or speech and noise masks ($K = 2$) on VBD

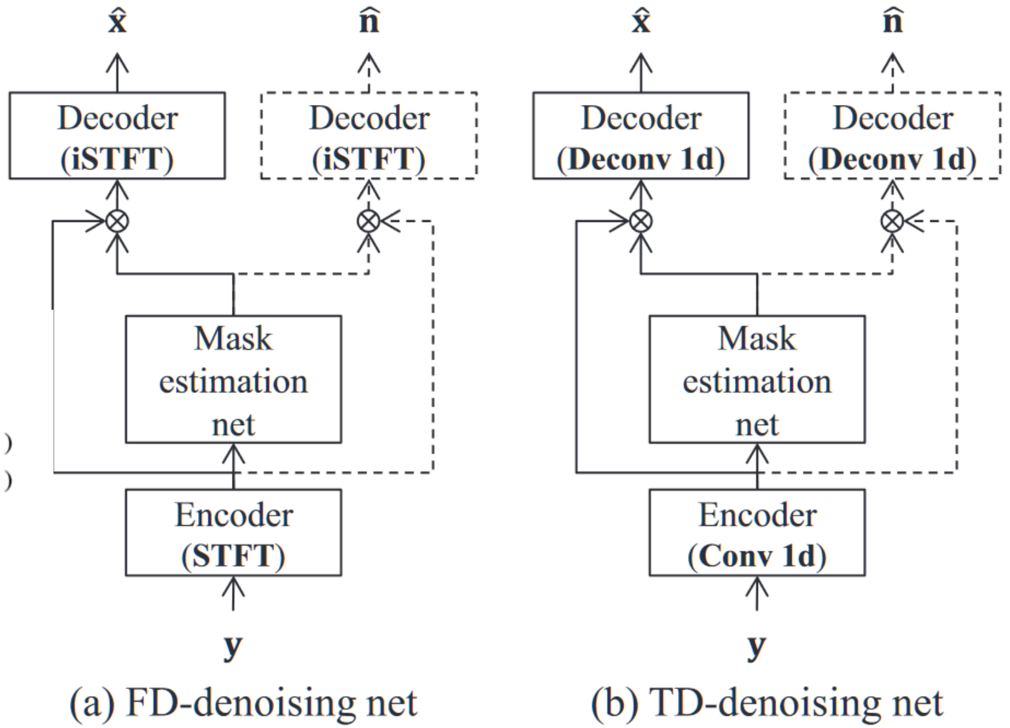


Figure from [1]

[1] Kinoshita, Keisuke, et al. "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network." ICASSP 2020.
[2] Koyama, Yuichiro, et al. "Exploring the Best Loss Function for DNN-Based Low-latency Speech Enhancement with Temporal Convolutional Networks." *arXiv preprint arXiv:2005.11611* (2020).

Trend #5: From Research to Deployment

- Recent progress allows commercial use of audio source separation
  - Quality reached level that was unthinkable five years ago
  - Computational resources on edge devices sufficient for real-time inference (e.g., smartphone, tinyML)

- Example: Real-time Speech enhancement
  - separate noisy speech into speech and noise – many demos on YouTube

Google Meet noise cancellation

NVIDIA RTX Voice

krisp.ai

Trend #5: From Research to Deployment

- Example: Karaoke
  - Traditionally: Karaoke uses “cover-version” material split into vocals/accomp.
    - Not all songs available for Karaoke
    - Varying degree of quality of Karaoke songs
  - Now: Using source separation allows to enjoy Karaoke version of any song
    - Spotify SingAlong [1]
      - Allows to turn down the volume of the vocals and shows synchronized lyrics
    - Line Music Japan/Taiwan [2]
      - Realizes Karaoke feature by on-device, real-time source separation from Sony
      - Allows to enjoy Karaoke version of many songs from the catalogue

Figure from [2]

[1] https://research.atspotify.com/making-sense-of-music-by-extracting-and-analyzing-individual-instruments-in-a-song/
[2] https://prtimes.jp/main/html/rd/p/000000004.000061313.html

Trend #5: From Research to Deployment

- Example: Sony Xperia 1 II "Intelligent Wind Filter"
  - Reduces wind noise for an even clearer audio recording
  - Demo: https://www.youtube.com/watch?v=tR_MHXpyIwA

- Example: Columbia Classics 4K Ultra HD Collection
  - Sony AI separation was used to create the Dolby ATMOS version of two movies ("Lawrence of Arabia" and "Gandhi")

More Open Problems

- Speech enhancement w/ sample rate of 48kHz lacks "ecosystem"
  - No large-scale, clean studio-quality data available
    - VBD small, MCV noisy + bandlimited
  - Perceptual measures (PESQ, ⋯) are only there for 8kHz or 16kHz
- Phase reconstruction for music separation
  - Many works in this direction [1, 2, ⋯] but not really yet a breakthrough
  - This is in contrast to speech separation/enhancement tasks
- Better use of transformers
  - Transformers dominate sequence modeling in NLP
  - However, in source separation (not yet) so commonly used
  - Interesting idea: Complex transformer which was proposed in [3]?

[1] Takahashi, Naoya, et al. "PhaseNet: Discretized Phase Modeling with Deep Neural Networks for Audio Source Separation." InterSpeech 2018.
[2] Le Roux, Jonathan, et al. "Phasebook and friends: Leveraging discrete representations for source separation." JSTSP 2019
[3] Yang, Muqiao, et al. "Complex Transformer: A Framework for Modeling Complex-Valued Sequence." ICASSP 2020

Contents

- Introduction to Audio Source Separation
- Current Trends and Open Problems
- Ecosystems, Datasets and Upcoming Competitions

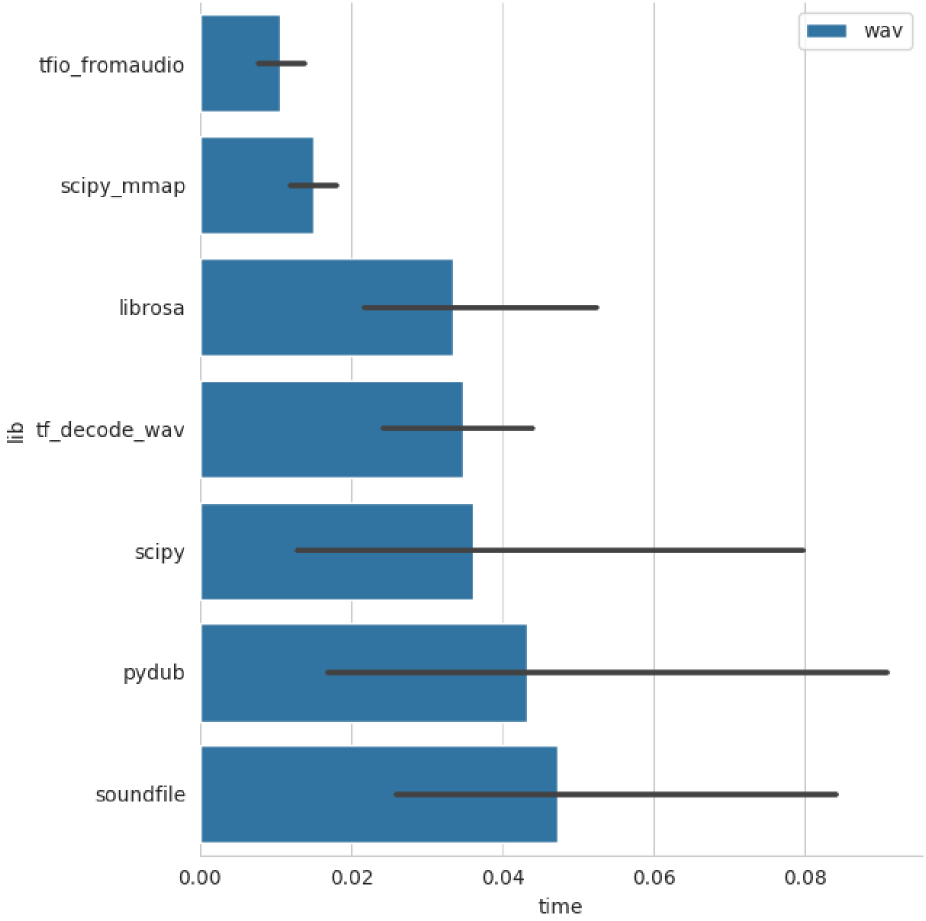
Ecosystems for Audio Source Separation

The importance of open source for research

- Music Separation 2019
  - State of the art: SONY corporation systems presented ICASSP 2015
  - Open (and popular) implementations 2.5 dB behind state of the art!
  - In 2019 a many (pretrained) Open Source systems were released
    - Open-Unmix (2019) `sigsep/open-unmix-pytorch`
    - Demucs (2019) `facebookresearch/demucs`
    - Spleeter (2019) `deezer/spleeter`
    - Nussl (2020) `nussl/nussl`
- Speech Separation
  - Same trend but new systems are all open source
    - Asteroid (2020) `mpariente/asteroid`

Other audio tools

- Sampling strategies, how to chunk/batch
  - Scaper `justinsalamon/scaper`
  - MUDA `bmcfee/muda`
  - Wavaugment `facebookresearch/WavAugment`
- Move pre-processing to GPU
  - DDSP `magenta/ddsp`
  - Torchaudio `pytorch/audio`
  - Karpe
- Source separation often has I/O bottlenecks
  - tfio `tensorflow/io`
  - Dali `NVIDIA/DALI`



https://github.com/faroit/awesome-python-scientific-audio

https://github.com/faroit/python_audio_loading_benchmark/

Datasets for Audio Source Separation

Single Utterance / Tracks

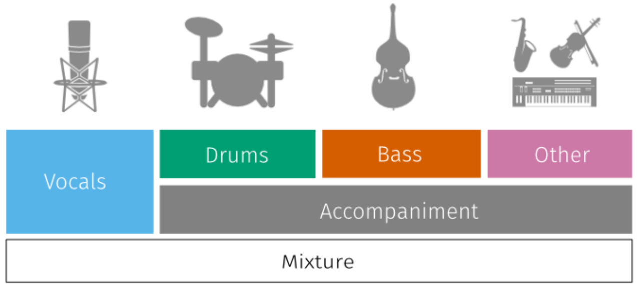
Music Separation:
- MedleyDB
- Slakh

Speech
- WSJ
- VCTK
- VoiceBank+Demand
- DNS challenge dataset
- Mozilla commonvoice

Supervised Separation

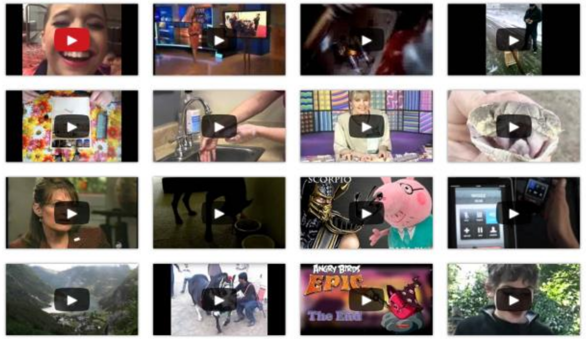
Music Separation:
- MUSDB18

Speech
- WSJ0X–mix
- Librimix
- WHAM/WHAMR



Universal Separation

- AudioSet
- WildMix
- Free Sound Dataset
- …



Upcoming Competitions

- SiSEC 2021
  - Music Separation Challenge
  - For the first time private test data will be used specifically produced for challenge

- Deep Noise Suppression Challenge
  - First challenge is/was together with InterSpeech 2020
  - Second challenge has started and will be run together with ICASSP 2021
  - More information: https://dns-challenge.azurewebsites.net

Conclusions

- Is the problem solved?
  - "A system that achieves human auditory analysis performance in all listening situation" (Wang)
  - Thanks to deep learning a lot of progress has been made
  - But from the contest results we know that we are not there yet

- Resources – Overview Papers

  - Wang, DeLiang, and Jitong Chen. "Supervised speech separation based on deep learning: An overview." IEEE/ACM Transactions on Audio, Speech, and Language Processing 26.10 (2018): 1702-1726
  - Rafii, Zafar, et al. "An overview of lead and accompaniment separation in music." IEEE/ACM Transactions on Audio, Speech, and Language Processing 26.8 (2018): 1307-1335.
  - Gannot, Sharon, et al. "A consolidated perspective on multimicrophone speech enhancement and source separation." IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.4 (2017): 692-730

Thank you for your attention

If you have any questions, then please contact us

fabian-robert.stoter@inria.fr

stefan.uhlich@sony.com