Target speech extraction

Marc Delcroix

Collaborators


T. Ochiai


K. Zmolikova[1]


K. Kinoshita


H. Sato


J. Bennasar Vázquez


A. Ogawa


N. Tawara


T. Nakatani


S. Araki


Y. Ohishi

[1]Brno University of Technology

Cocktail party-effect

Humans can focus their attention intentionally on a specific sound signal (Selective hearing)

Realized using various clues
[Darwin+00]

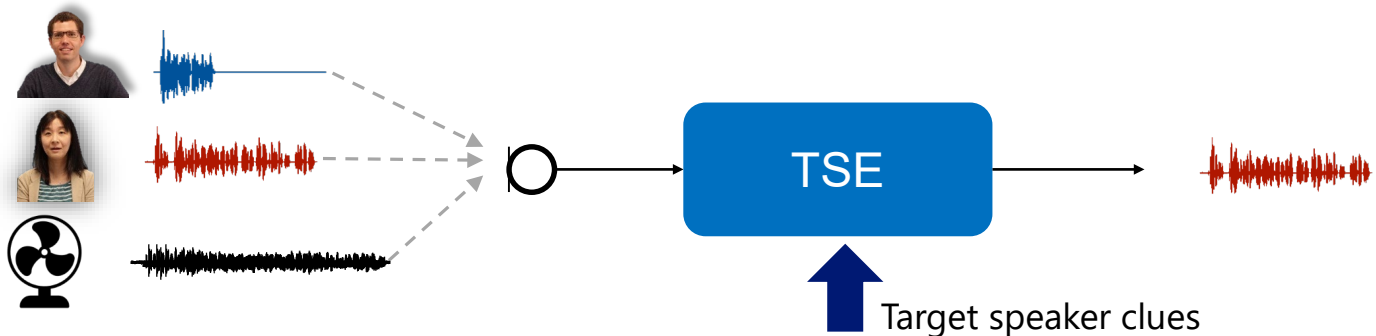
- Locational,
- Speaker voice characteristics,
- Visual,

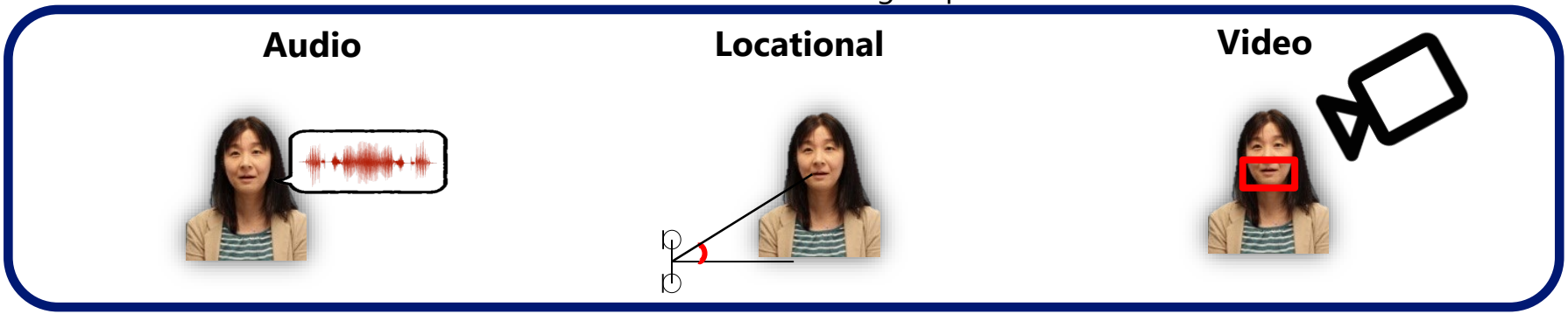
→ Can follow a conversation at a cocktail party

Target speech extraction (TSE)

Computational selective hearing

i.e. Extract speech of a target speaker in a mixture given speaker clues

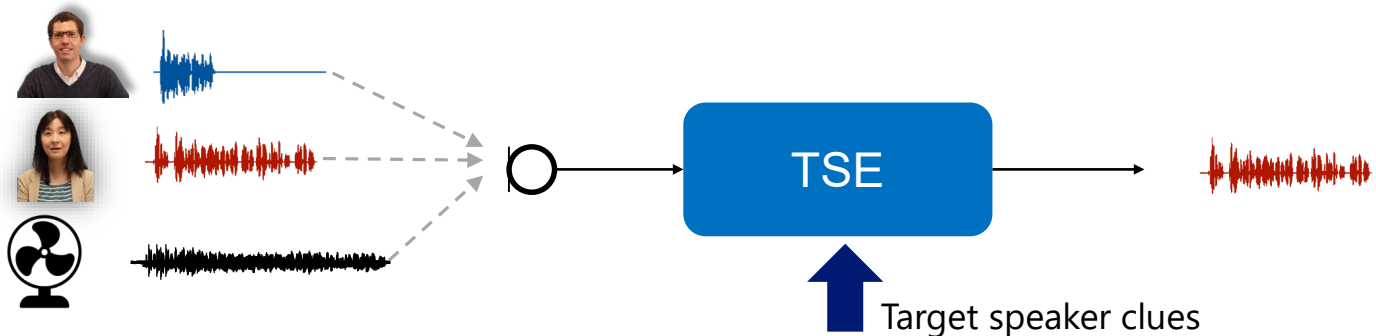


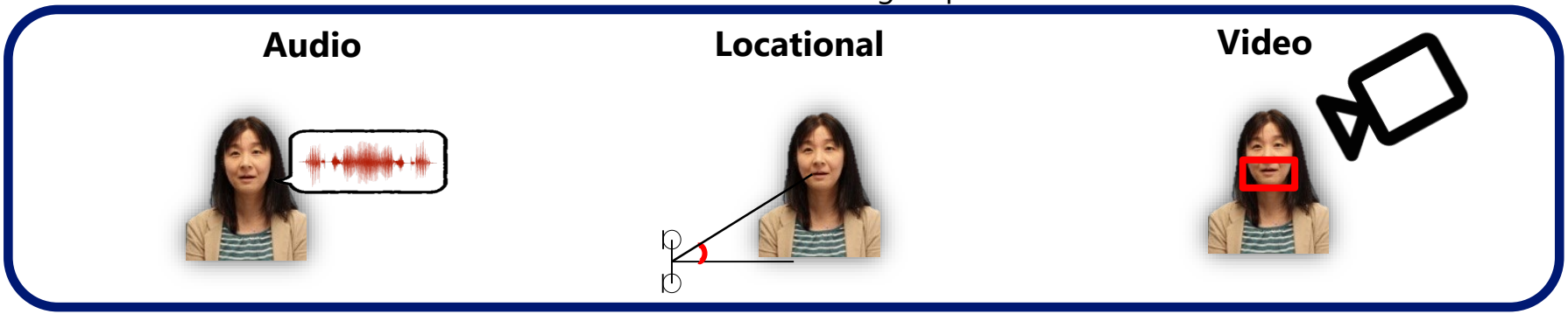


Target speech extraction (TSE)

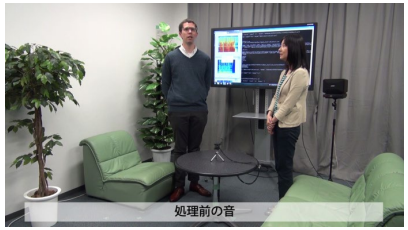
Computational selective hearing

i.e. Extract speech of a target speaker in a mixture given speaker clues





Demo video



Demo of audio-clue-based TSE (SpeakerBeam)

(1) Record enrollment

(2) TSE demo: man + woman

Available on YouTube:
https://www.youtube.com/watch?v=7FSHgKip6vI

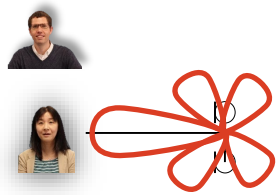
Classical approaches

Fixed beamformer

- Extract signal from a fixed direction

☹ Requires knowing the position of the target speaker

→ Lack of flexibility



Separation

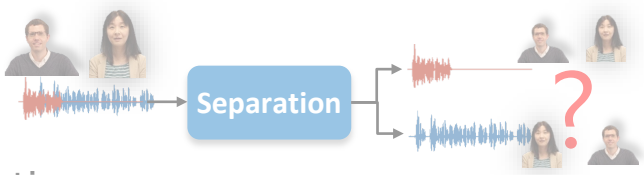
- Separate mixture into all its source signals

☹ Requires knowing/estimating number of speakers

☹ Speaker-output ambiguity

→ Need to be combined with some speaker identification

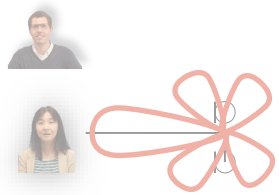
→ Cascade separation+ speaker identification is not optimal for TSE



Classical approaches

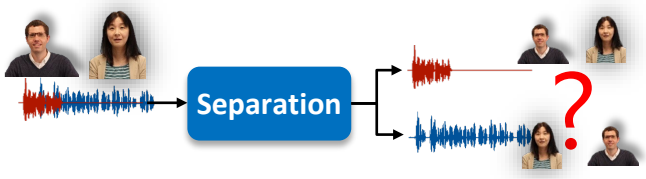
Fixed beamformer

- Extract signal from a fixed direction
- ☹ Requires knowing the position of the target speaker
- → Lack of flexibility



Separation

- Separate mixture into all its source signals
- ☹ Requires knowing/estimating number of speakers
- ☹ Speaker-output ambiguity
  - → Need to be combined with speaker identification
  - → Cascade separation+ speaker identification is not optimal for TSE



Advantages of TSE

By exploiting speaker clues, TSE avoids the limitations of previous schemes

- ☺ **No need to know the speaker location**
- ☺ **No need to know the number of speakers**
- ☺ **No speaker-output ambiguity**
- ☺ **Optimal**

TSE made possible recently thanks to progress in speech enhancement/separation, speaker identification

Especially, deep-learning enabled optimized TSE systems

- 2017, showed possibility with **audio clues** [Zmolikova17]
- 2018, showed possibility with **video clues** [Afouras+18, Ephrat+18,Owens +18]

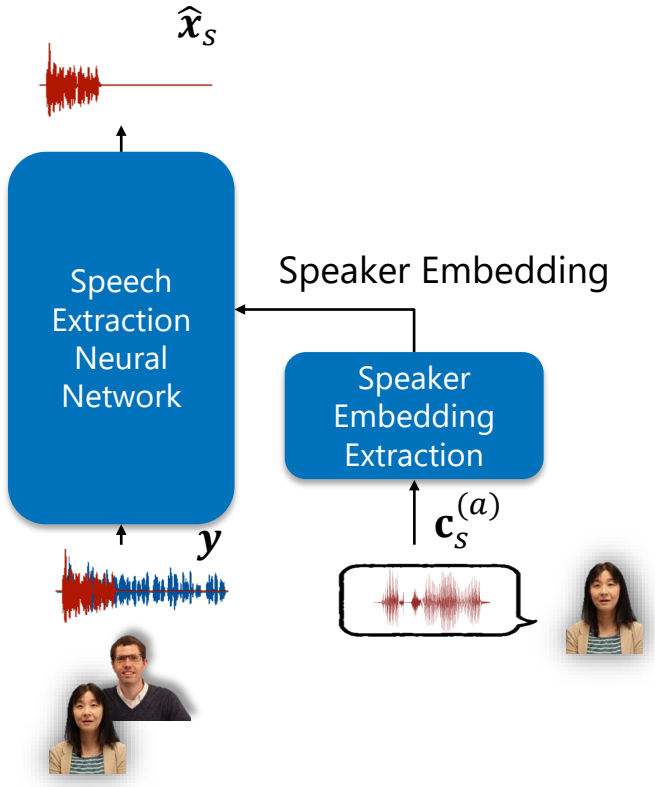
  Since then,

- Rapid progress following development of neural speech enhancement/separation/speaker identification

Audio-clue-based extraction TSE

Use an enrollment utterance of the target speaker (few seconds of audio) to inform which voice to extract in the mixture [Zmolikov+17]

→ Speech separation & Speaker identification at once

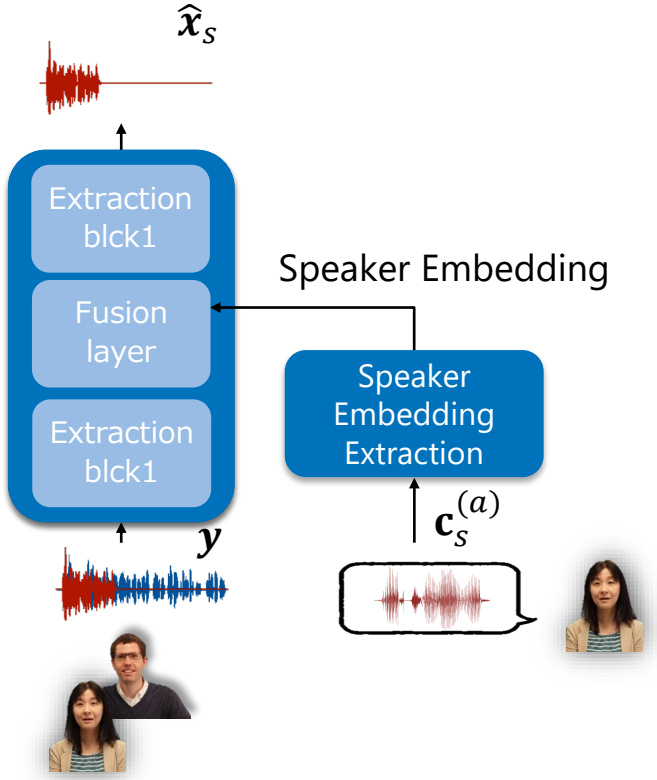


Audio-clue-based extraction TSE

Use an enrollment utterance of the target speaker (few seconds of audio) to inform which voice to extract in the mixture [Zmolikov+17]

→ Speech separation & Speaker identification at once

- Various ways to implement audio-clue-based TSE
  - What speaker embeddings?
    - i-vectors, d-vectors [Wang+19], jointly-learned [Zmolikova+17b]
  - What type of extraction networks?
    - Fusion layer: Concatenation[Wang+19, Xu+19], Multiplication[Delcroix+19a], Factorized layer [Zmolikova+17a], Attention [Xiao+19, Li+19]
    - BLSTM, CNN etc,
    - Frequency domain, Time-domain
    - Regression, Mask

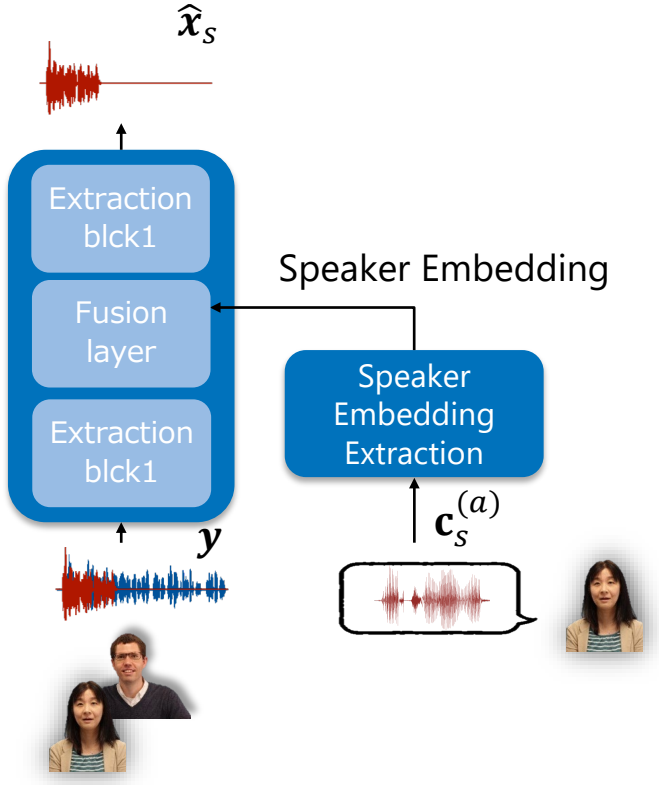


Audio-clue-based extraction TSE

Use an enrollment utterance of the target speaker (few seconds of audio) to inform which voice to extract in the mixture [Zmolikov+17]

→ Speech separation & Speaker identification at once

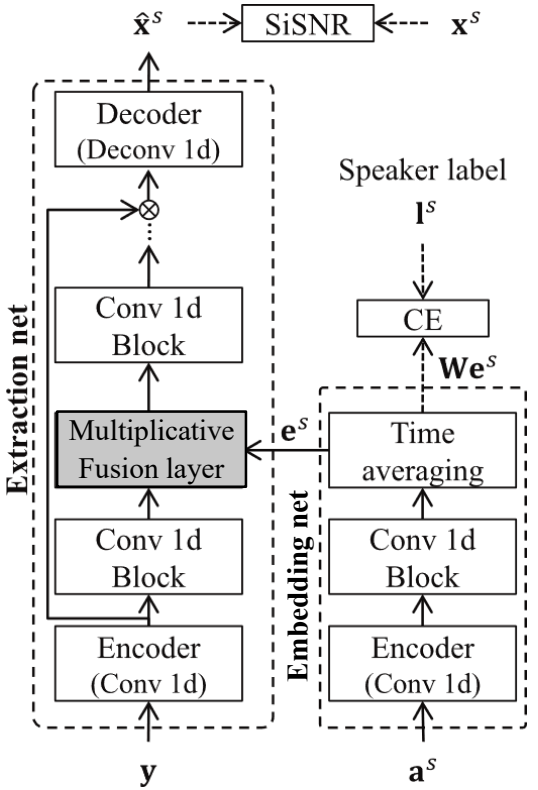
- Various ways to implement audio-clue-based TSE
  - What speaker embeddings?
    - i-vectors, d-vectors [Wang+19], **jointly-learned** [Zmolikova+17b]
  - How to integrate embeddings (fusion layer)?
    - Concatenation[Wang+19, Xu+19], **Multiplication**[Delcroix+19a], Factorized layer [Zmolikova+17a], Attention [Xiao+19, Li+19]
  - What type of extraction networks
    - BLSTM, **CNN** etc,
    - Frequency domain, **Time-domain**
    - Regression, **Mask**



SpeakerBeam

Delcroix et al. “Improving speaker discrimination of target speech extraction with time-domain SpeakerBeam,” ICASSP, 2020.

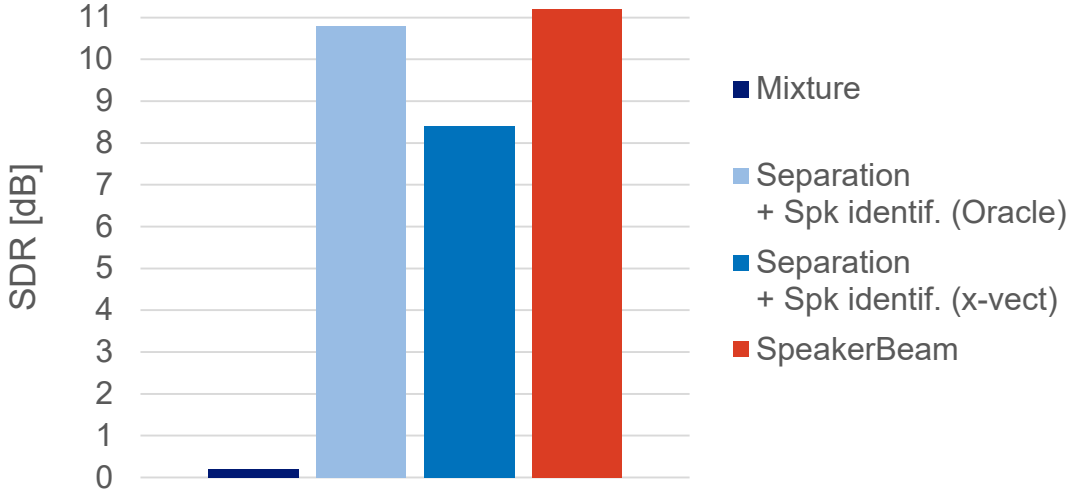
- Time-domain SpeakerBeam

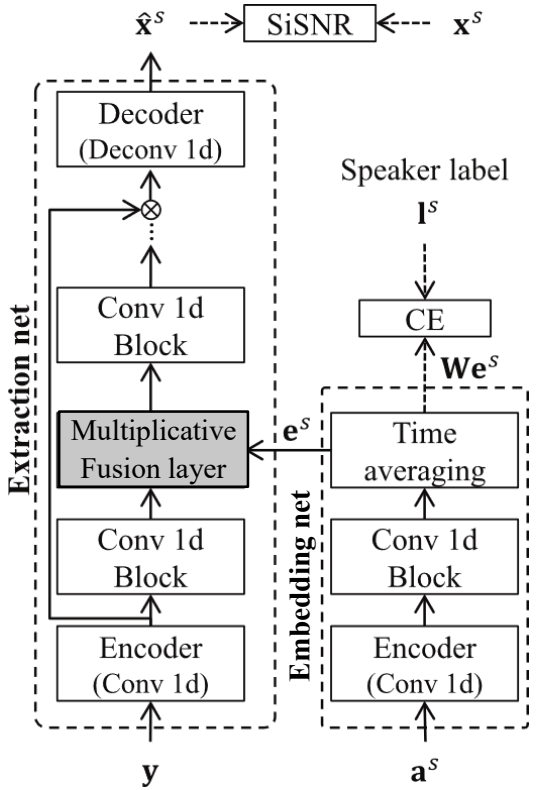


SpeakerBeam

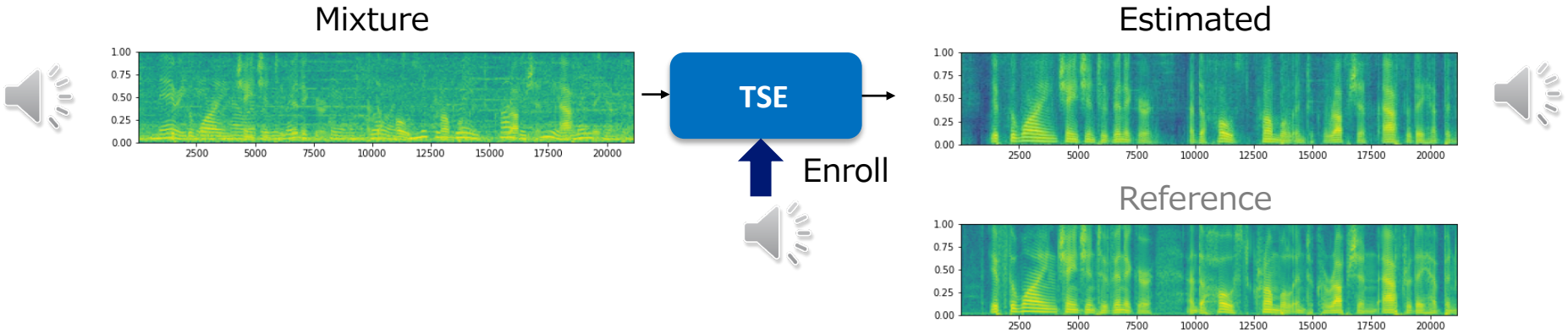
Delcroix et al. “Improving speaker discrimination of target speech extraction with time-domain SpeakerBeam,” ICASSP, 2020.

- Time-domain SpeakerBeam
- Tested on MC-WSJ-2mix (reverberant)
- Evaluation metric: Signal-to-distortion (SDR) [dB]

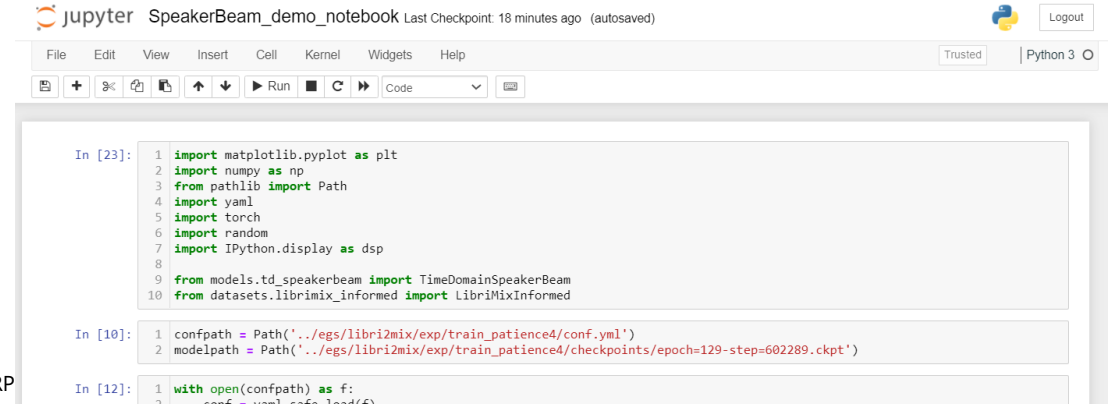




Sound demo



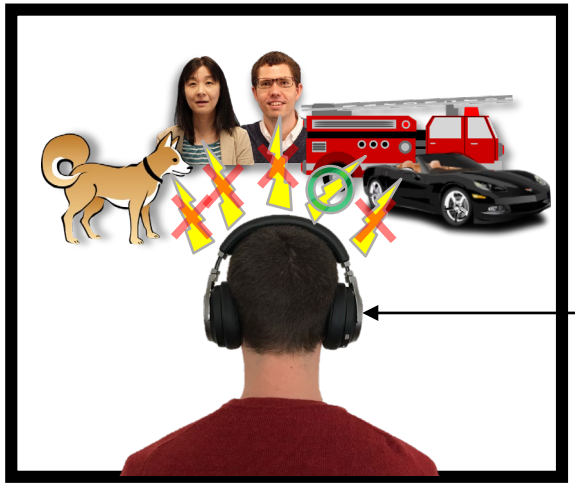
Code: https://github.com/BUTSpeechFIT/speakerbeam



Extension to target sound extraction

Extension to non-speech signals

→ Extract target *sound* from a *mixture of sounds*

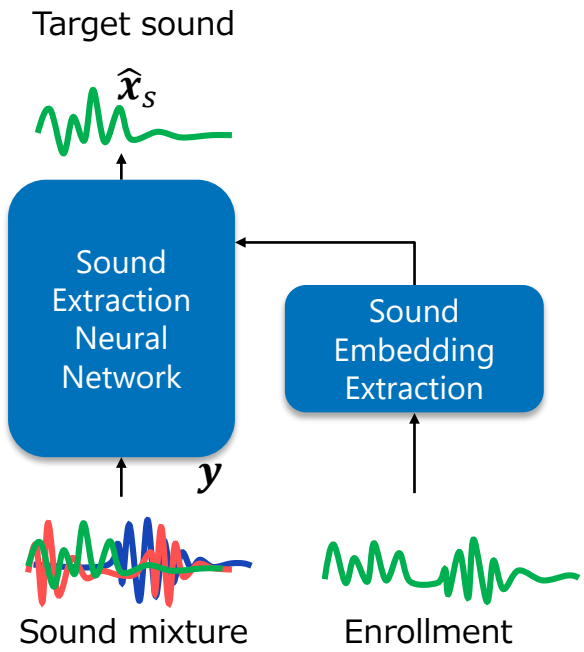


Target sounds	
Dog	×
Woman	×
Man	×
Siren	○
Car	×

2 ways to realize

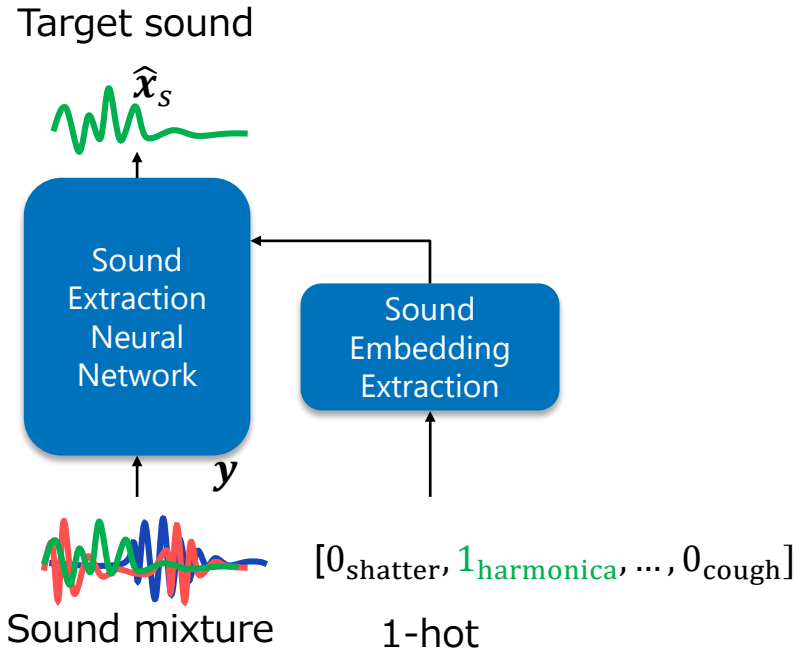
Enrollment-based approach

[Zmolikova+17, Lee+19, Gfeller+21]



1-hot-based approach

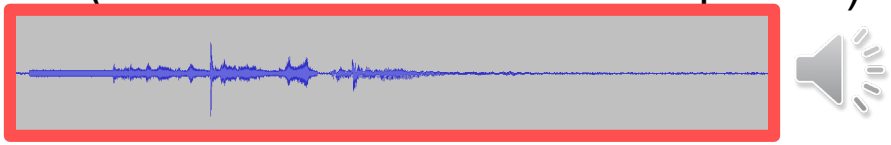
[Ochiai+20, Kong+20]



+ Combination of both: Delcroix et al., *"Few-Shot Learning of New Sound Classes for Target Sound Extraction," Interspeech 2021*

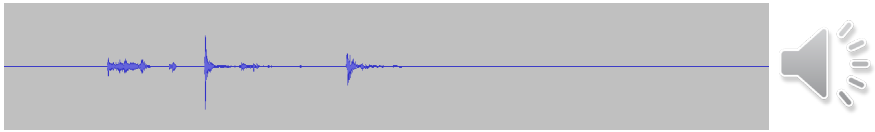
Sound demo

Mixture (“Harmonica”+ “Shatter” + “Speech”)

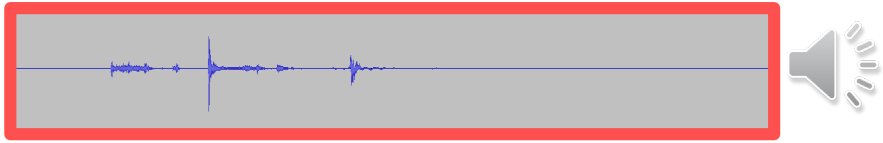


Seen sound class

Target (“Shatter”)



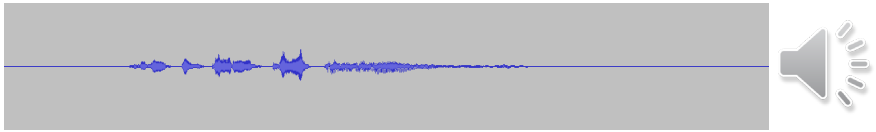
Extracted (“Shatter”)



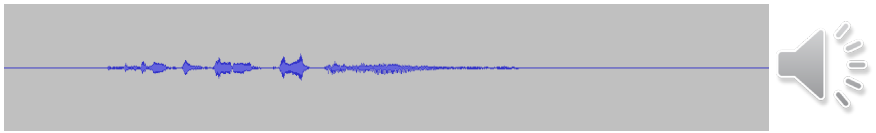
SDR improvement: 12.92 dB

New sound class

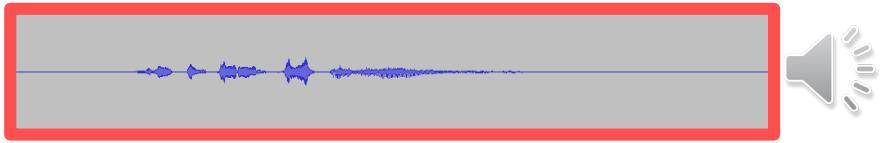
Target (“Speech”)



Extracted w/o retraining (“Speech”)



Extracted w/ retraining (“Speech”)



SDR improvement: 13.54 dB

Conclusion

- Target speech/sound extraction is a promising way to tackle the cocktail party problem
- Idea can be applied to other problems:
  - Audio-visual extraction [Afouras+18, Ephrat+18]
  - Target speaker ASR [King+17, Delcroix+18, Kanda+19, Delcroix+19, Denisov+19]
  - Target speaker VAD [Ding+20, Medennikov+20]
  - EEG-based attentive listening [O'Sullivan+14, Aroudi+20]
- To dig further:
  - Demo video of SpeakerBeam: https://www.youtube.com/watch?v=7FSHgKip6vI
  - Slide of Interspeech tutorial: https://butspeechfit.github.io/tse_tutorial
  - SpeakerBeam implementation: https://github.com/BUTSpeechFIT/speakerbeam

Thank you!

Questions?

Email me at marc.delcroix@ieee.org