Neural Target Speech and Sound Extraction

Marc Delcroix

November 20th, 2025

Agenda

Introduction

Part 1: Target speech extraction

- Separation vs TSE
- Origin of TSE & SpeakerBeam
- Diffusion-based TSE

Part 2: Target sound extraction

- Class-label vs Enrollment-based approaches
- SoundBeam
- Continuous learning

Main collaborators

Including interns, current and past colleagues


Katerina Zmolikova


Tsubasa Ochiai


Keisuke Kinoshita


Hiroshi Sato


Jorge Bennasar Vázquez


Junyi Peng


Ján Švec


Carlos Hernandez-Olivan


Y. Ohishi


Naoyuki Kamo


Takafumi Moriya


Atsunori Ogawa


Naohiro Tawara

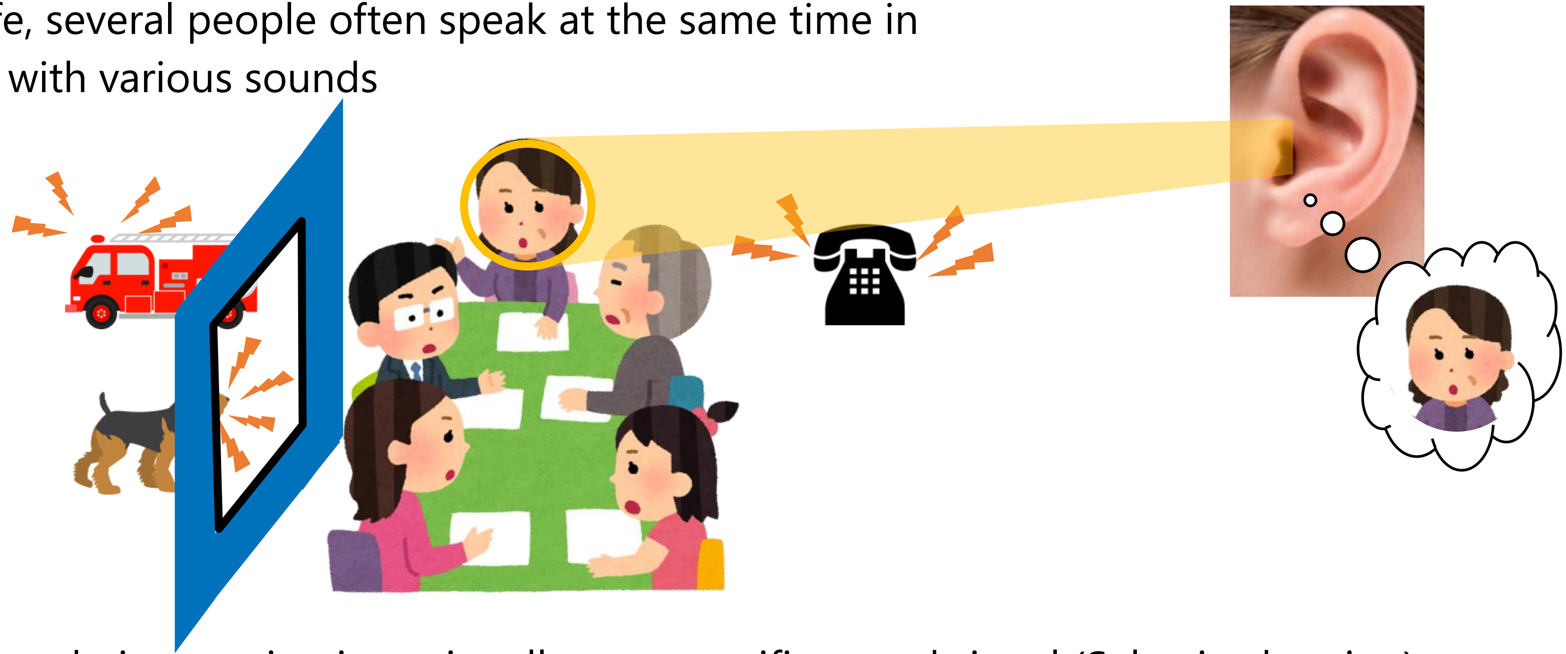

Tomohiro Nakatani


Shoko Araki


Honza Černocký

Selective hearing

In everyday life, several people often speak at the same time in environments with various sounds



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

Selective hearing

In everyday life, several people often speak at the same time in environments with various sounds

Humans can focus their attention intentionally on a specific sound signal (Selective hearing)
Realized using various clues, e.g., locational, speaker voice/sound characteristics, visual, semantic

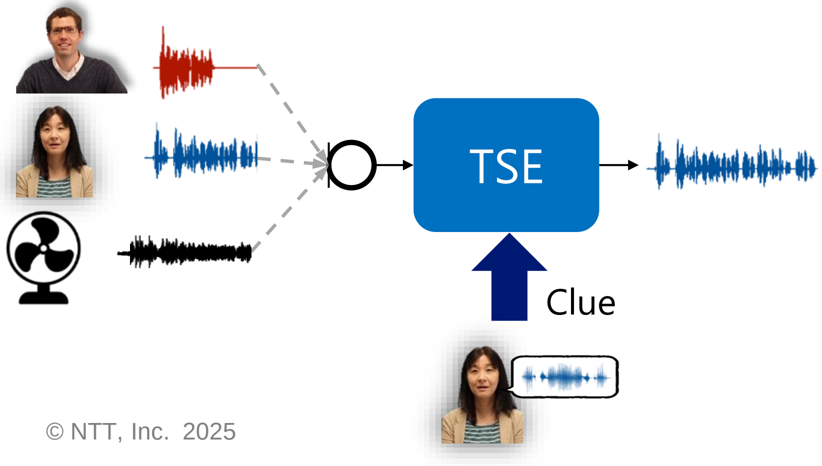
→ It allows us to follow a conversation at a cocktail party, pick up our name, etc.

Target speech/sound extraction (TSE)

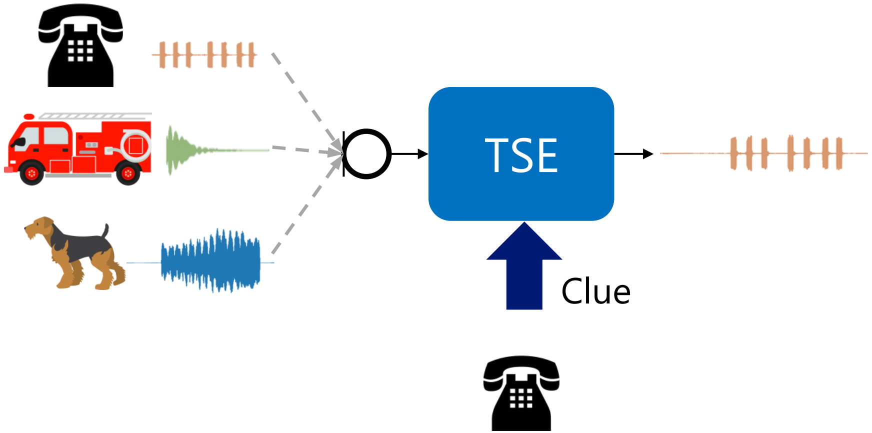
Goal: realize computational selective hearing

Extract signal of a target speaker or desired sound in a mixture, given clues about the target

Target speech extraction



Target sound extraction



Wide range of possible applications

Speech enhancement (SE)

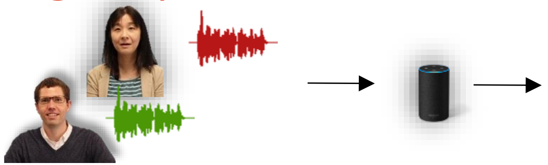
- Hearing aids/hearables
- Teleconference
- Voice recorder

Wide range of possible applications

Speech enhancement (SE)

- Hearing aids/hearables
- Teleconference
- Voice recorder

Target-Speaker ASR



Hello, what's the weather today?

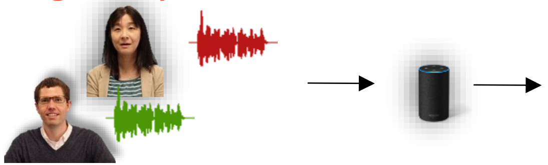
- Smartphones
- Smart speakers

Wide range of possible applications

Speech enhancement (SE)

- Hearing aids/hearables
- Teleconference
- Voice recorder

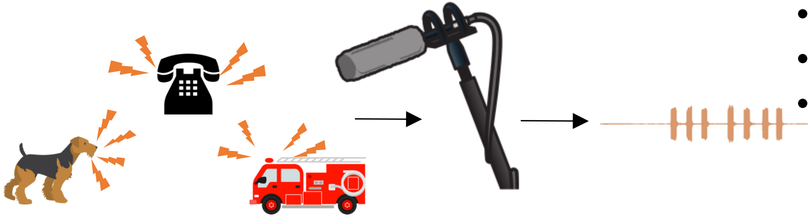
Target-Speaker ASR



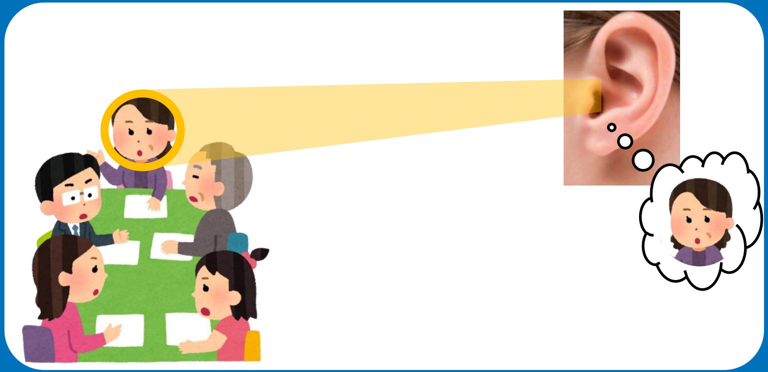
Hello, what's the weather today?

- Smartphones
- Smart speakers

Sound extraction



- Sound post-production
- Remixing
- hearables



Target Speech extraction

Agenda

Introduction

Part 1: Target speech extraction

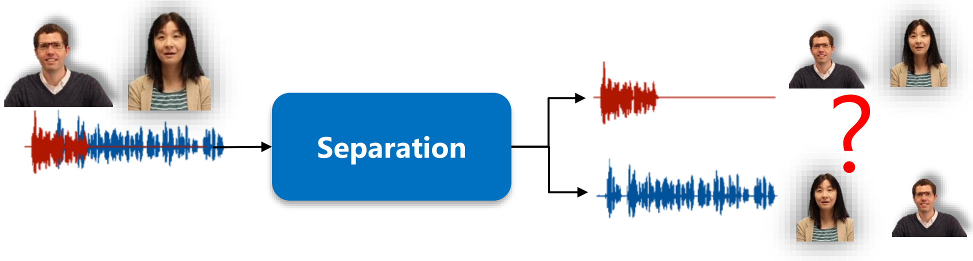
- Separation vs TSE
- Origin of TSE & SpeakerBeam
- Diffusion-based TSE

Part 2: Target sound extraction

- Class-label vs Enrollment-based approaches
- SoundBeam
- Continuous learning

Separation

Separate mixture into all its source signals



☹ Requires knowing/estimating number of sources

☹ Source-output ambiguity

→ Need to be combined with speaker identification, which may cause error propagation

TSE

Extract only the target speaker

→ *Speech separation & speaker identification at once*

By exploiting target clues, TSE avoids the limitation of separation schemes

☺No estimation of the number of sources required

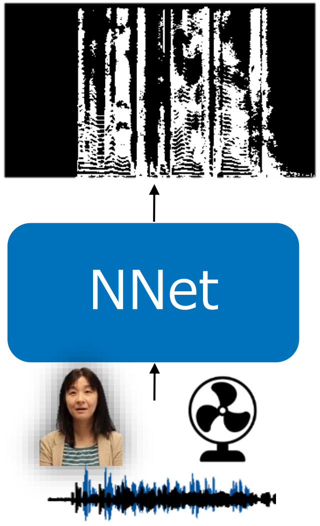
☺No source-output ambiguity

Origin of neural TSE ~ 2016

Origin of TSE ~2016: DNN-based noise reduction

[Hori'15, Heymann’15]

Speech enhancement with DNN-based TF-mask estimation



Speech + Noise

TF-mask

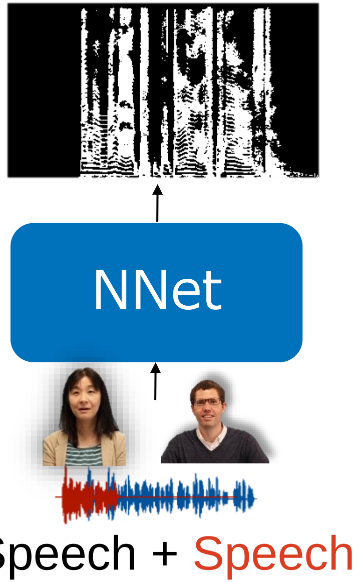
$$M(t,f) = \begin{cases} 1 & \text{if speech} > noise \\ 0 & \text{otherwise} \end{cases}$$

Extract speech from noise

- Can train a DNN to discriminate speech from noise using simulated mixtures
- Promissing approach especially when combined with beamforming (e.g., CHiME3 [Heymann’15])

Origin of TSE ~2016: Speaker dependent extraction

Speech extraction with DNN-based TF-mask estimation



TF-mask

$$M(t,f) = \begin{cases} 1 & \text{if target speech} > \text{interference} + \text{noise} \\ 0 & \text{otherwise} \end{cases}$$

→How to identify the target?

Solution at the time:

- Speaker dependent (always same target!)
- Dominance-based, gender-based

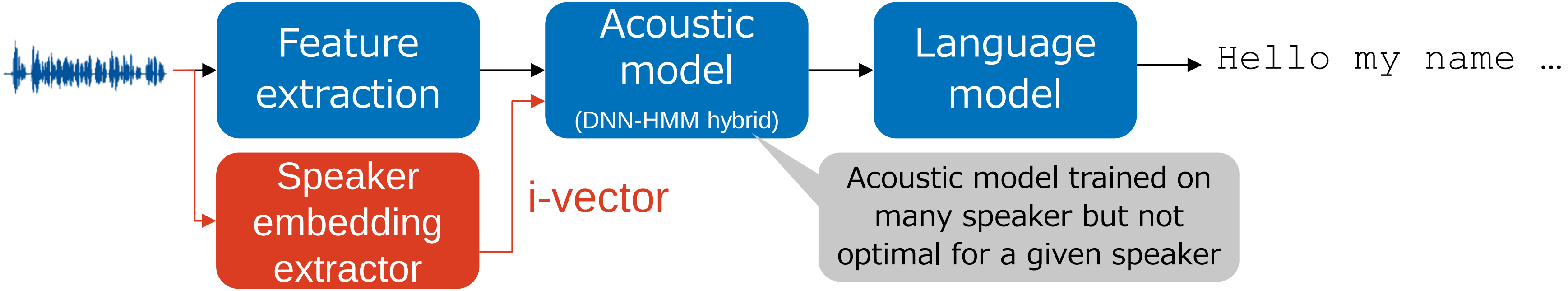
☹ **Do not generalize well!**

Could we do some kind of model adaptation instead?

Origin of TSE ~2016: Acoustic model adaptation

[Saon'13, Delcroix'15]

Speaker adaptation for ASR (before end-to-end ASR)

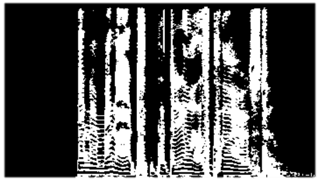


Speaker adaptation with i-vectors was popular at the time
- Adapting a general acoustic model to specific speaker characteristics
- Dealt with single speaker (**not speech mixtures**)

Origin of TSE ~2016: Target speech extraction

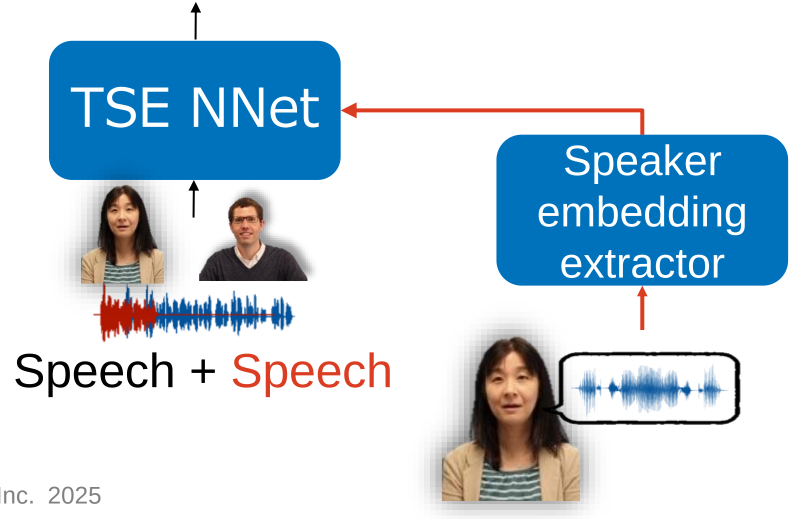
[Zmolikova'17]

Use speaker adaptation idea to inform the NNet about the target speaker



TF-mask

$$M(t,f) = \begin{cases} 1 & \text{if target speech} > \text{interference} + \text{noise} \\ 0 & \text{otherwise} \end{cases}$$



Extract speaker embedding from an enrollment utterance:
about 10 sec. of speech from the target speaker

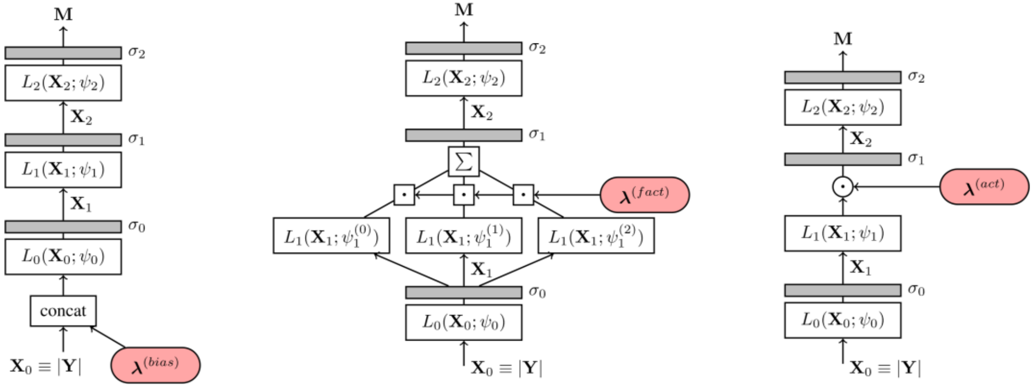
→ Much more flexible!

Initial investigations: SpeakerBeam

[Zmolikova'17, Zmolikova'19]

Conditioning

- Input bias
- Factorized layer
- **Multiplication**
- (Addition, FiLM, Attention, etc.)



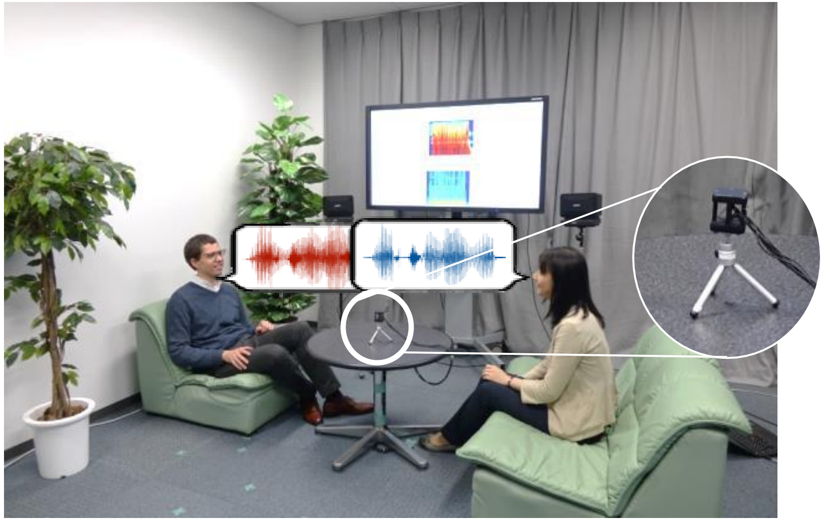
Speaker representation

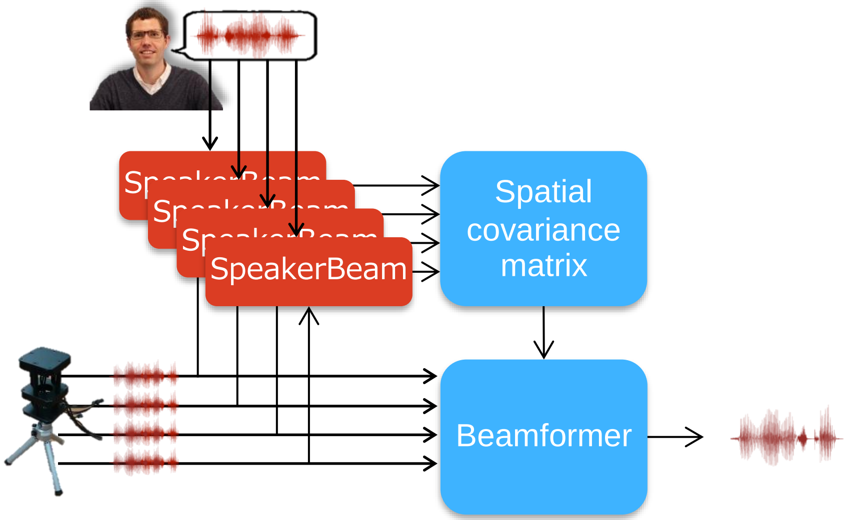
- Speaker posterior (Train w/ 1-hot, test with posteriors)
  - › *Poor for same-sex mixtures of unseen speakers*
- i-vector
- **Jointly-learned embedding extractor**

Conditioning	Speaker representation	SDRi [dB] ↑
Input bias	i-vector	-3.8
	Jointly learned	-2.2
Factorized	i-vector	5.7
	Jointly learned	**6.1**
Multiplicative	i-vector	5.2
	Jointly learned	5.6

WSJ-2mix

Demo video





Watch the demo video on YouTube:
https://www.youtube.com/watch?v=7FSHgKip6vI

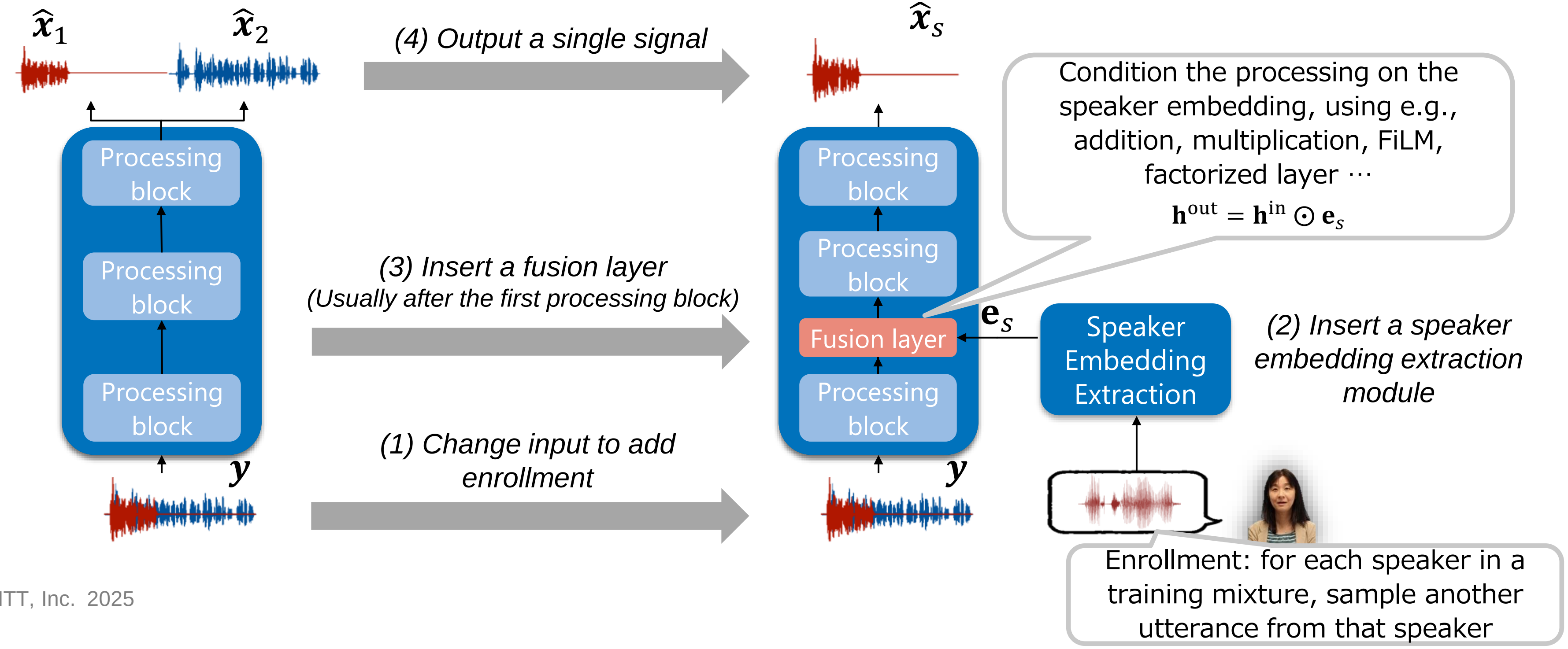
How to build a TSE system?

From speech separation to TSE

$$\mathcal{L}^{PIT} = \min \begin{pmatrix} \mathcal{L}(\hat{\boldsymbol{x}}_1(\boldsymbol{\theta}), \boldsymbol{x}_1) + \mathcal{L}(\hat{\boldsymbol{x}}_2(\boldsymbol{\theta}), \boldsymbol{x}_2) \\ \mathcal{L}(\hat{\boldsymbol{x}}_1(\boldsymbol{\theta}), \boldsymbol{x}_2) + \mathcal{L}(\hat{\boldsymbol{x}}_2(\boldsymbol{\theta}), \boldsymbol{x}_1) \end{pmatrix}$$

(5) Remove PIT in the loss

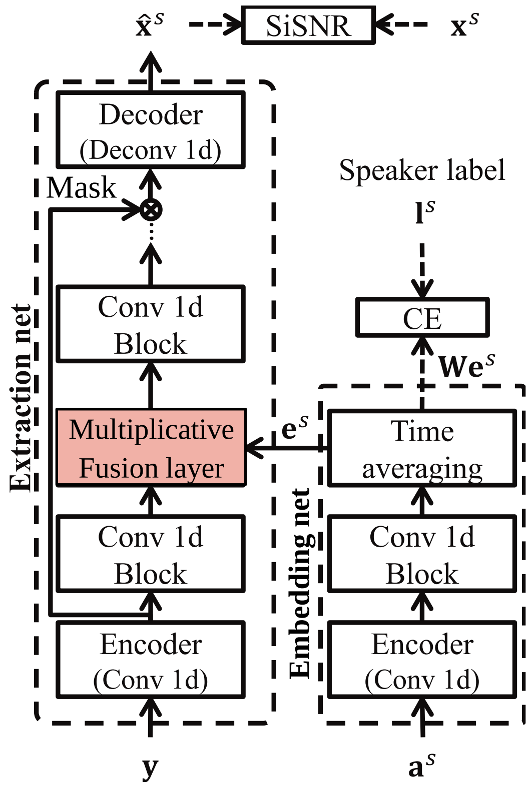
$$\mathcal{L}^{TSE} = \mathcal{L}(\hat{\boldsymbol{x}}_s(\boldsymbol{\theta}), \boldsymbol{x}_s)$$



Time-domain SpeakerBeam

[Delcroix+20]

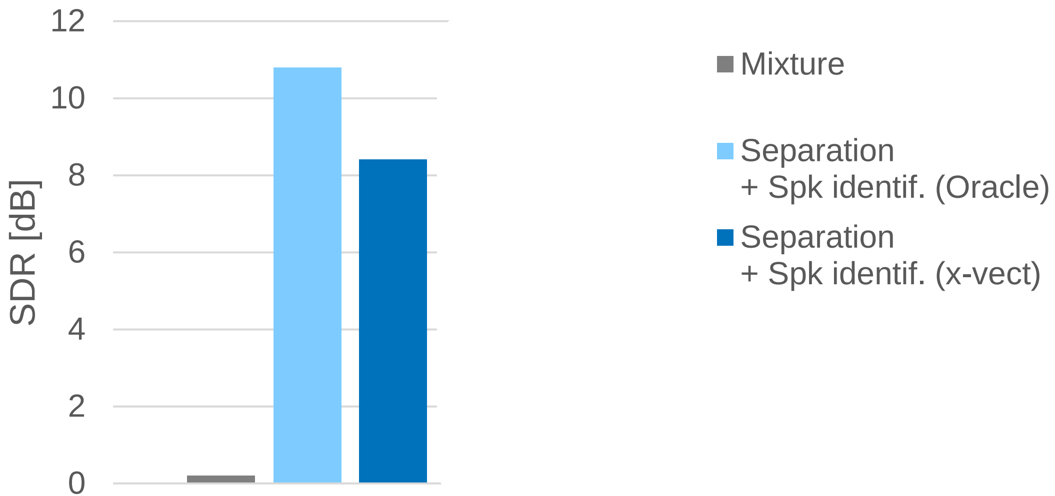
- Based on Conv-TasNet architecture [Luo+18]
- Tested on MC-WSJ-2mix (reverberant)
- Evaluation metric: Signal-to-distortion (SDR) [dB]

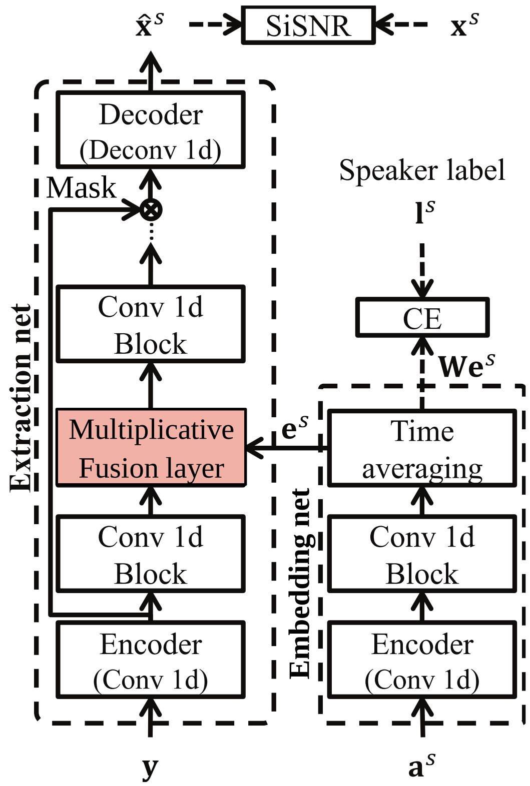


Time-domain SpeakerBeam

[Delcroix+20]

- Based on Conv-TasNet architecture [Luo+18]
- Tested on MC-WSJ-2mix (reverberant)
- Evaluation metric: Signal-to-distortion (SDR) [dB]

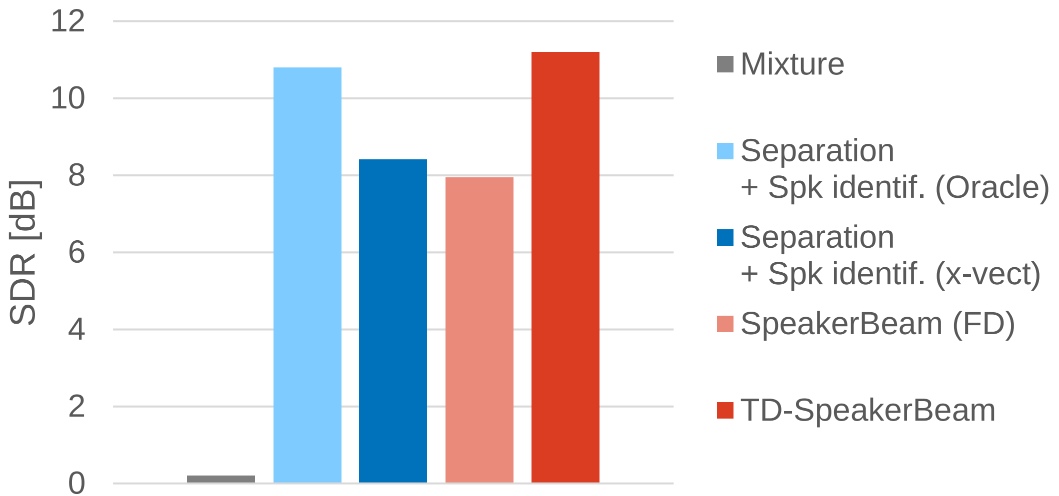


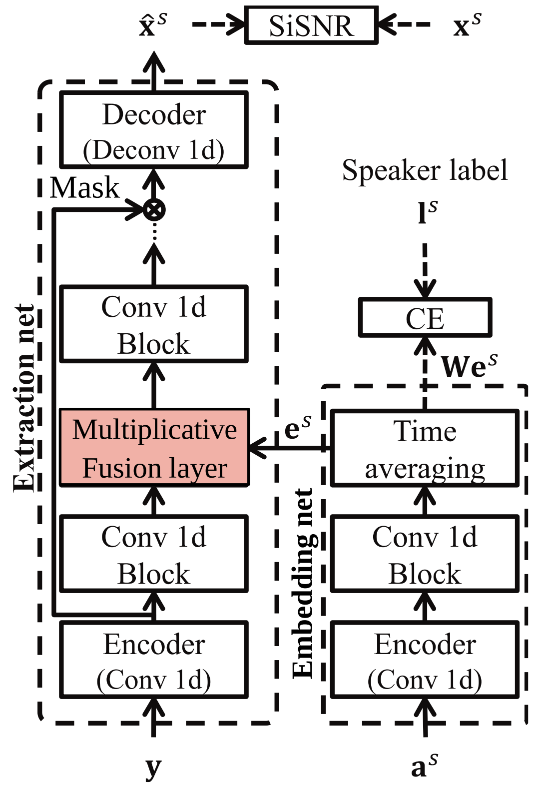


Time-domain SpeakerBeam

[Delcroix+20]

- Based on Conv-TasNet architecture [Luo+18]
- Tested on MC-WSJ-2mix (reverberant)
- Evaluation metric: Signal-to-distortion (SDR) [dB]





More natural TSE with diffusion model

[Kamo+23]

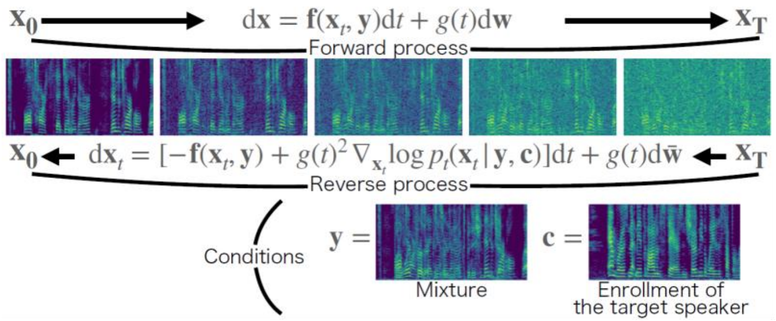
NTT

Exploit deep generative models, e.g., diffusion models

- Conventional conditional diffusion model models clean speech distribution [Welker’22] [Richter’23]

- Add condition on the target speaker clue $\mathbf{c}$:
  → Can perform TSE

$$p(\underbrace{\mathbf{x}_0}_{\text{Target speech}} \,|\, \underbrace{\mathbf{y}}_{\text{Mixture}}, \underbrace{\mathbf{c}}_{\text{Clue}})$$



N. Kamo et al., “Target Speech Extraction with Conditional Diffusion Model,” Interspeech 2023.

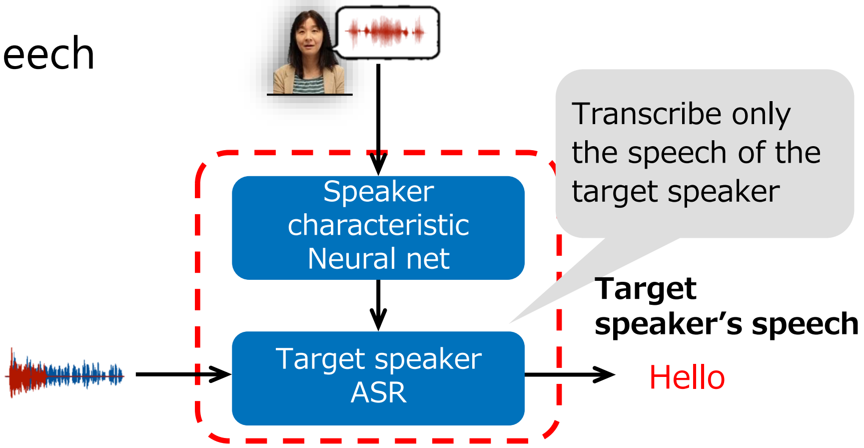
Wrap-up

- TSE can be realized with a neural network conditioned on speaker embedding
- Simple and practical idea
- Growing field
  - 2017: Initial ideas (SpeakerBeam)
  - 2018~19: Early follow-up works (Deep extractor network [Wang+18], VoiceFilter [Wang+19])
  - 2025: More than 30 related papers at ICASPP and Interspeech
- In parallel, other clues have been investigated
  - Visual clue-based TSE [Gabbay+18, Ephrat+18, Afouras+18, Ochiai+19]
  - EEG [Ceolini+20]
  - Speaker activity [Delcroix+21]
  - Semantic clues [Ohishi+22]

Application to other tasks

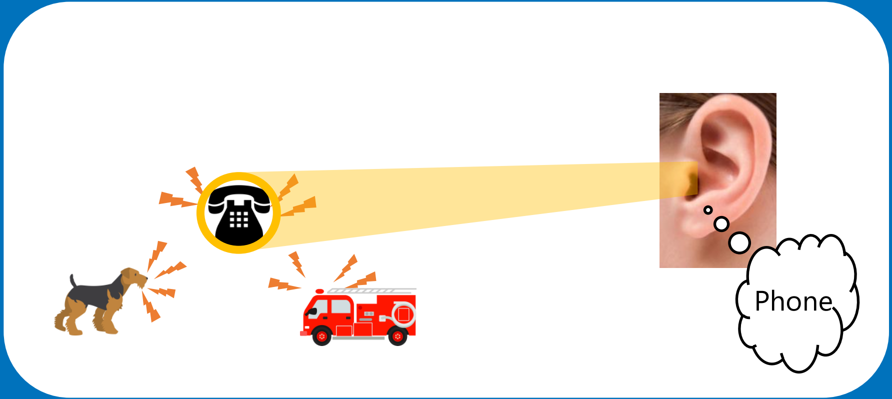
Same ideas can be applied to other tasks

- Target speaker ASR: directly transcribe speech of the target speaker without explicit extraction
  - DNN-HMM Hybrid [Delcroix+18, Zmolikova+18, Kanda+19]
  - Attention-based E2E systems [Delcroix+19 , Denisov+19, Shi+21]
  - Streaming system (RNN-T) [Moriya'22]



- Personalized VAD/ Target speaker VAD: find out when the target speaker is active [Sodoyer+06, Ding+20, Medennikov+20]

- Target sound extraction



Target Sound extraction

Agenda

Introduction

Part 1: Target speech extraction

- Separation vs TSE
- Origin of TSE & SpeakerBeam
- Diffusion-based TSE

Part 2: Target sound extraction

- Class-label vs Enrollment-based approaches
- SoundBeam
- Continuous learning

Target sound extraction

- We are often immersed in complex sound scenes with many sounds
  The same sound can be noise or carry important information depending on the situation

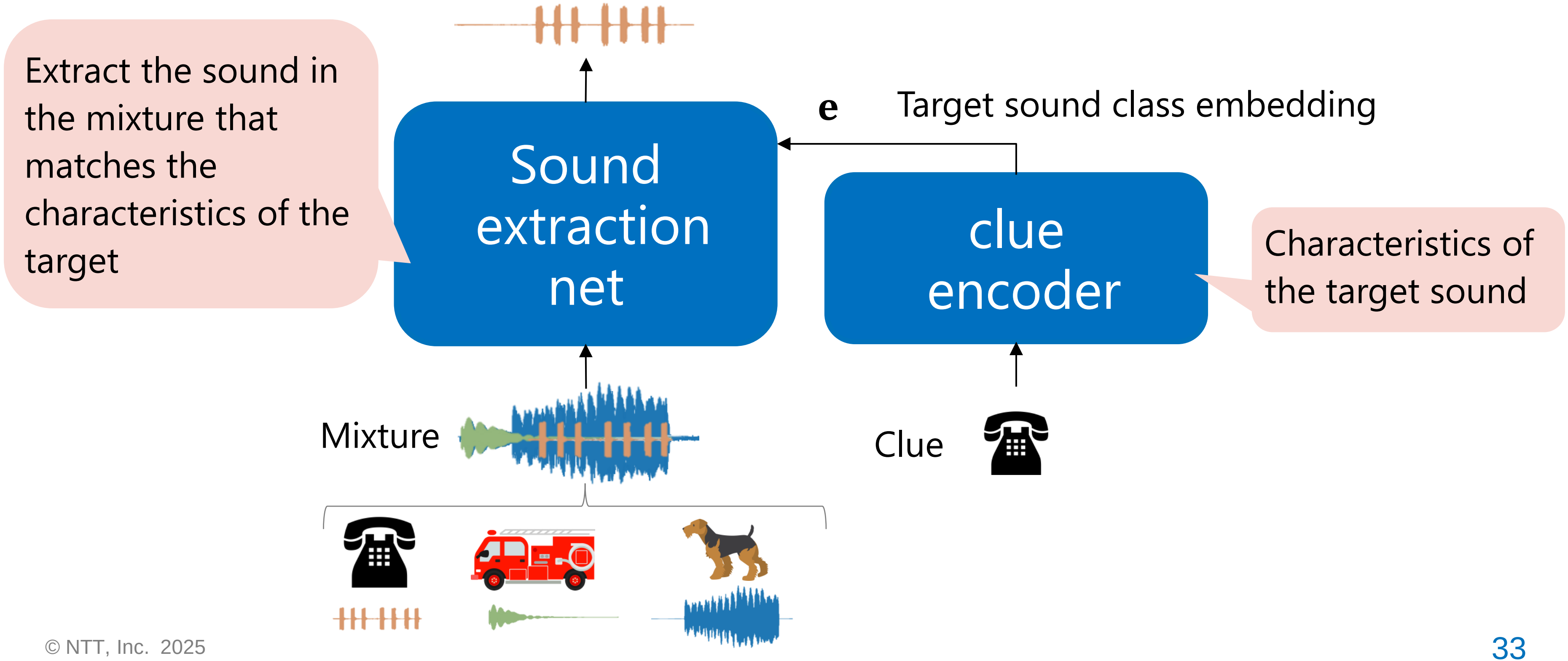
Siren when working at home
→ Noise

Siren when driving
→ Important sound

- Target sound extraction extends TSE (Speech) to arbitrary sounds
  - Sounds are more diverse → Easier to discriminate than speakers
  - Much more sound variety
    → Challenge to handle many sound classes
    → How to learn new sound classes (continuous learning)

Target sound extraction

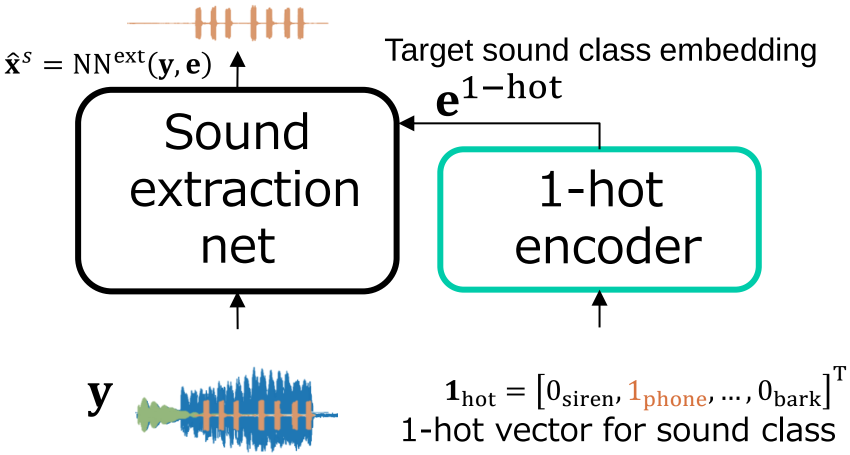
Can be realized with similar approach as target speech extraction



Sound-class vs Enrollment-based TSE

Sound class label-based TSE

[Ochiai+20, Kong+20]

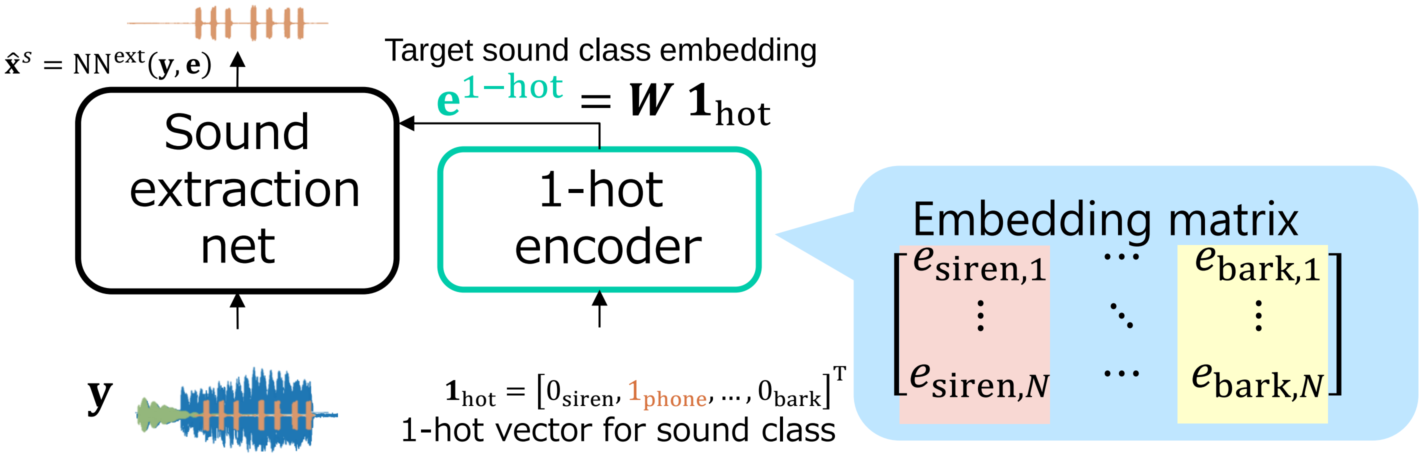


Sound-class vs Enrollment-based TSE

NTT

Sound class label-based TSE

[Ochiai+20, Kong+20]



Extraction via sound class labels

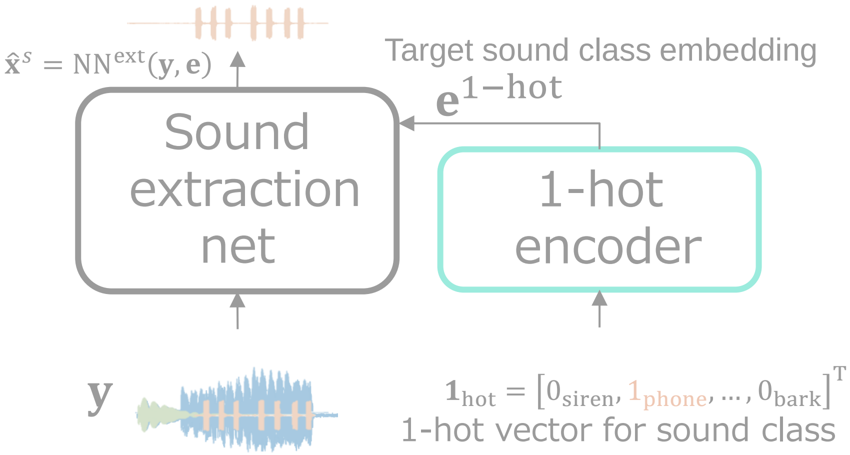
☺ Direct optimization of sound class embeddings

☹ Difficult to generalize to new sounds (unseen during training)

Sound-class vs Enrollment-based TSE

Sound class label-based TSE

[Ochiai+20, Kong+20]



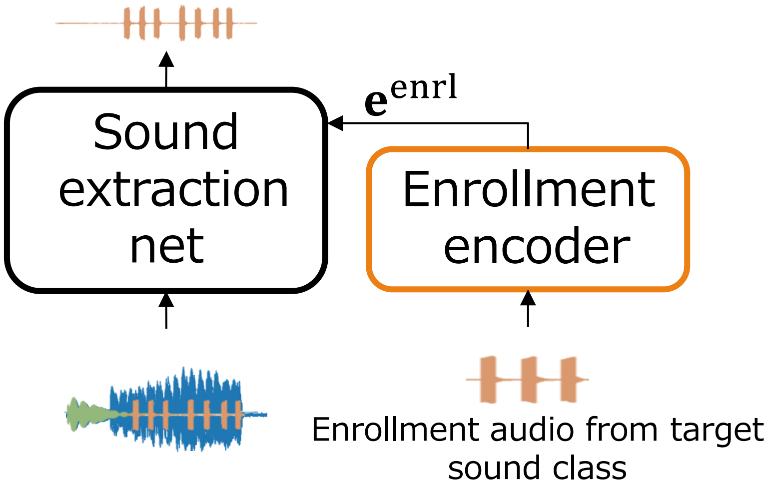
Extraction via sound class labels

☺ Direct optimization of sound class embeddings

☹ Difficult to generalize to new sounds (unseen during training)

Enrollment-based TSE

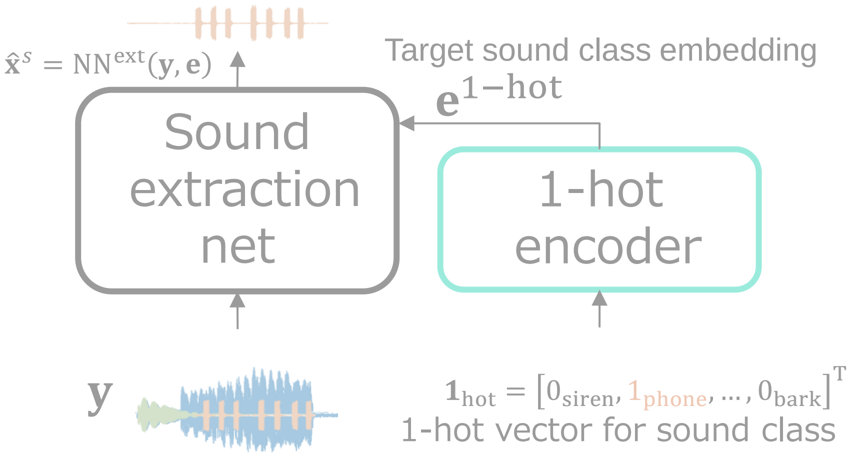
[Zmolikova+17 Lee+19, Gfeller+21]



Sound-class vs Enrollment-based TSE

Sound class label-based TSE

[Ochiai+20, Kong+20]



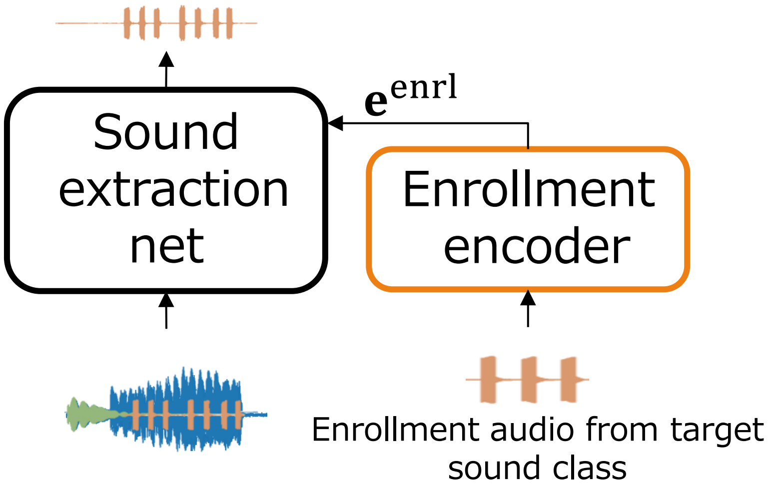
Extraction via sound class labels

☺ Direct optimization of sound class embeddings

☹ Difficult to generalize to new sounds (unseen during training)

Enrollment-based TSE

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



Extraction via sound similarity

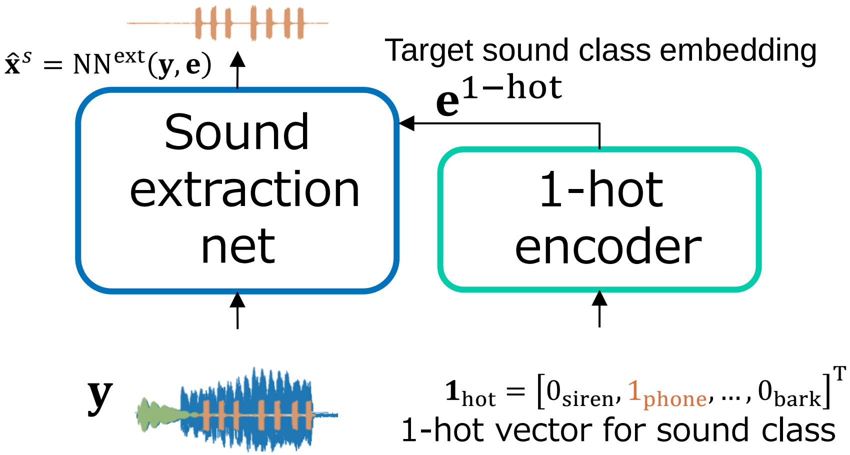
☺ Can generalize to new sounds

☹ Embeddings may not be optimal

Sound-class vs Enrollment-based TSE

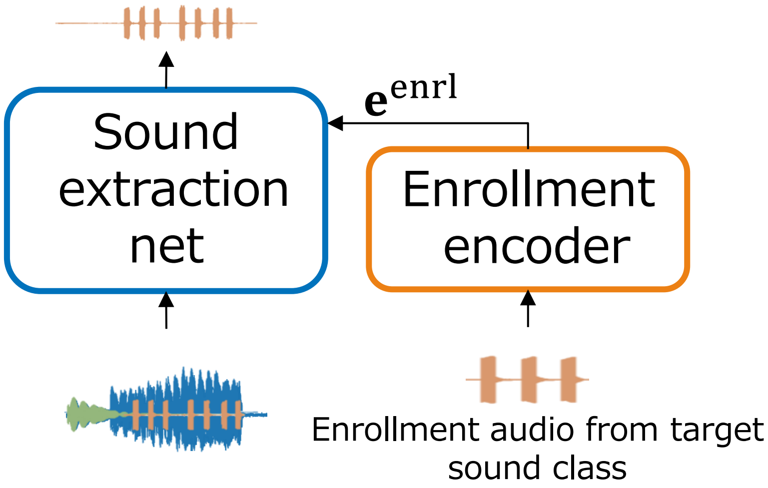
Sound class label-based TSE

[Ochiai+20, Kong+20]



Enrollment-based TSE

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



→ Can we combine the advantages of both frameworks?

Proposed mixed model: SoundBeam

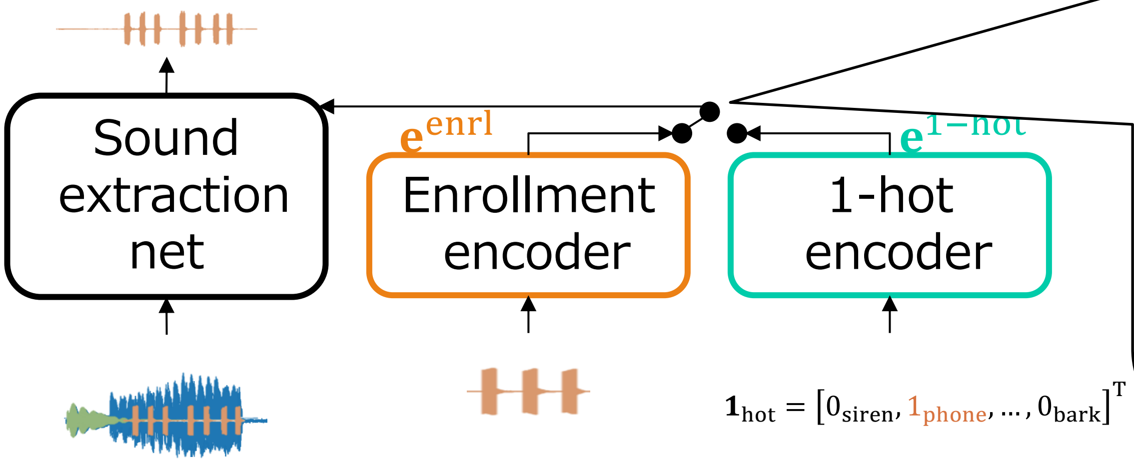
Combine both approaches: Jointly learning with 1-hot and enrollment with multi-task learning

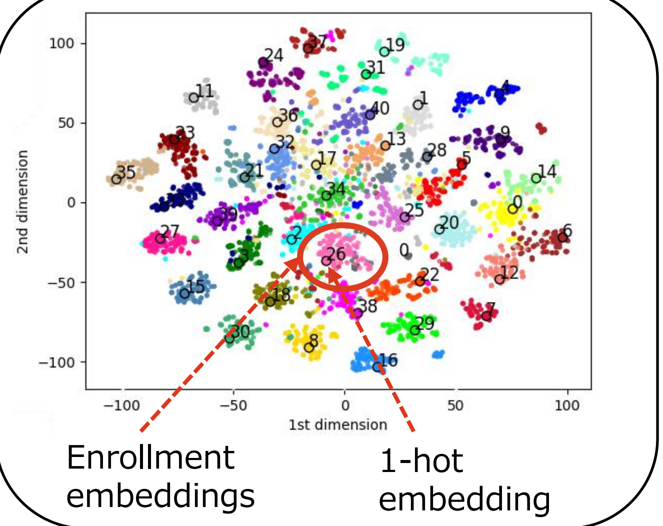
Isolated sound signal

Training loss $\mathcal{L}^{\text{SoundBeam}}(\hat{\mathbf{x}}^s, \mathbf{x}^s) = -\text{SNR}\left(\hat{\mathbf{x}}^s = \text{NN}^{\text{ext}}\left(\mathbf{y}, \mathbf{e}^{\text{enrl}}\right), \mathbf{x}^s\right) - \text{SNR}\left(\hat{\mathbf{x}}^s = \text{NN}^{\text{ext}}\left(\mathbf{y}, \mathbf{e}^{1-\text{hot}}\right), \mathbf{x}^s\right)$

→ Learn a common embedding space

$$\text{SNR} = 10\log\left(\frac{|\mathbf{x}^s|^2}{|\mathbf{x}^s - \hat{\mathbf{x}}^s|^2}\right)$$

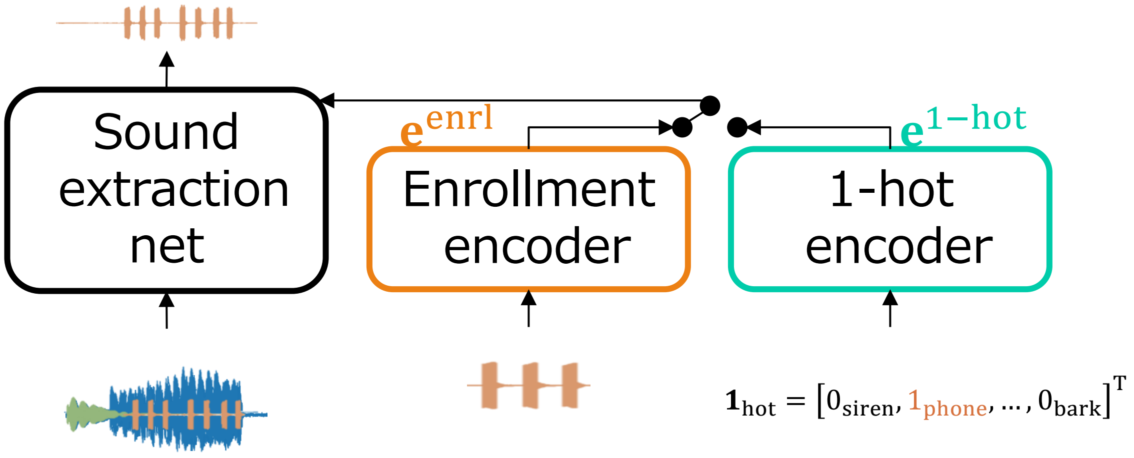




Proposed mixed model: SoundBeam

Combine both approaches: Jointly learning with 1-hot and enrollment with multi-task learning

- ☺ Improved performance thanks to multi-task learning
- ☺ High performance for seen classes
- ☺ Generalization to new sound classes



Adaptation to new sound classes

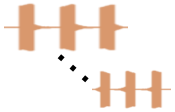
Goal: Add new sound classes to an existing TSE system

Using a few audio samples of cat sounds, enable cat sound extraction for a TSE system not trained to extract cat.

- Learn new entries to the 1-hot embedding matrix for the new sound class

$$W = \begin{bmatrix} e_{\mathrm{shatter},1} & \cdots & e_{\mathrm{cough},1} & e_{\mathrm{new},1} \\ \vdots & \ddots & \vdots & \vdots \\ e_{\mathrm{shatter},N} & \cdots & e_{\mathrm{cough},N} & e_{\mathrm{new},N} \end{bmatrix}$$

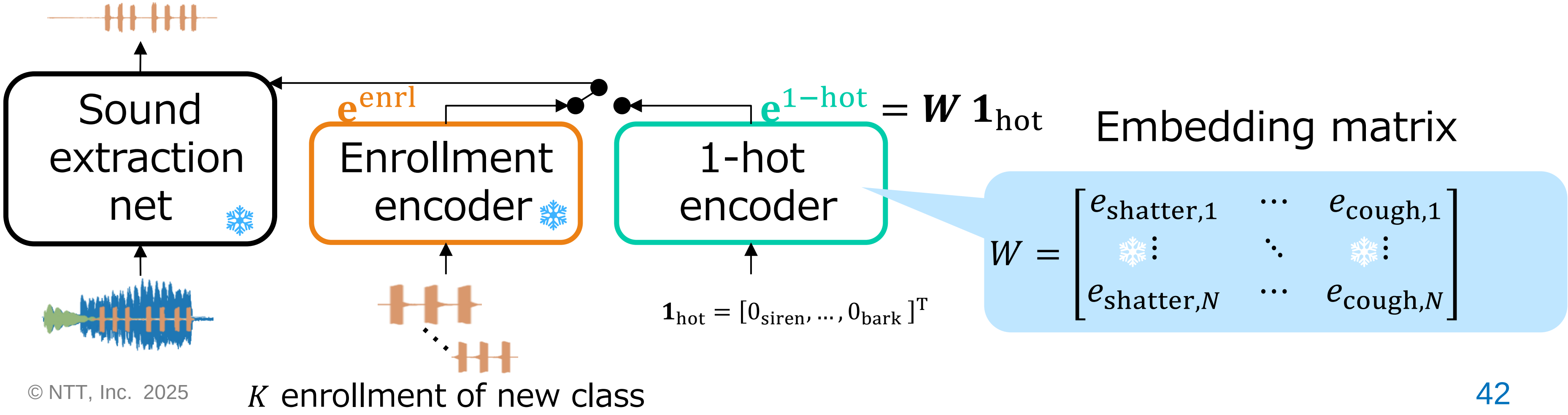
- Use only a few audio samples (e.g., K=5) from new sound class



- Do not modify the behavior for already learned classes

Adaptation to new sound classes

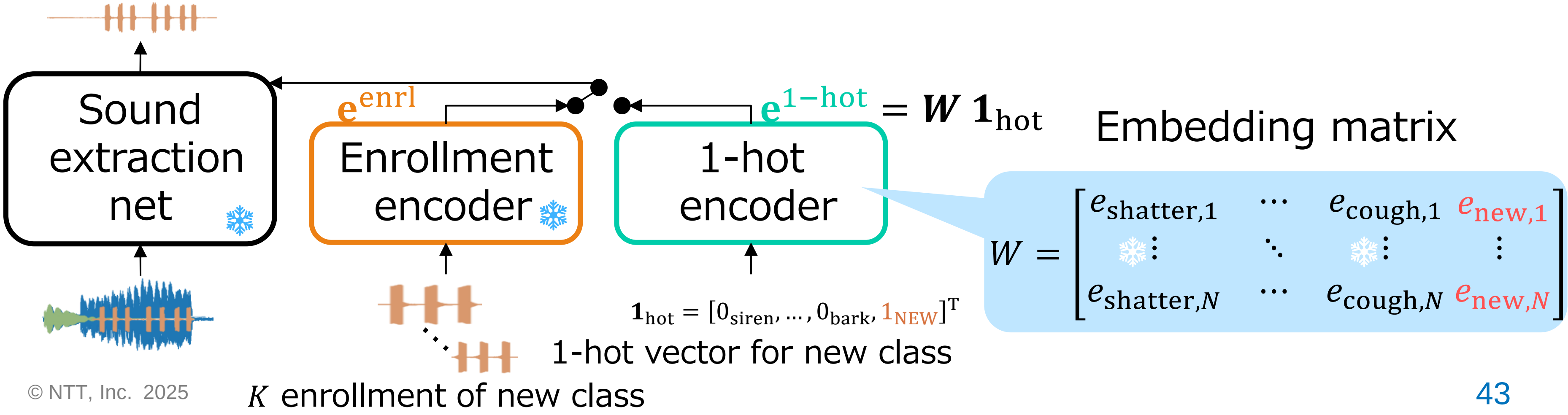
We assume a few enrollment samples (e.g., K=5) from new sound class



Adaptation to new sound classes

We assume a few enrollment samples (e.g., K=5) from new sound class

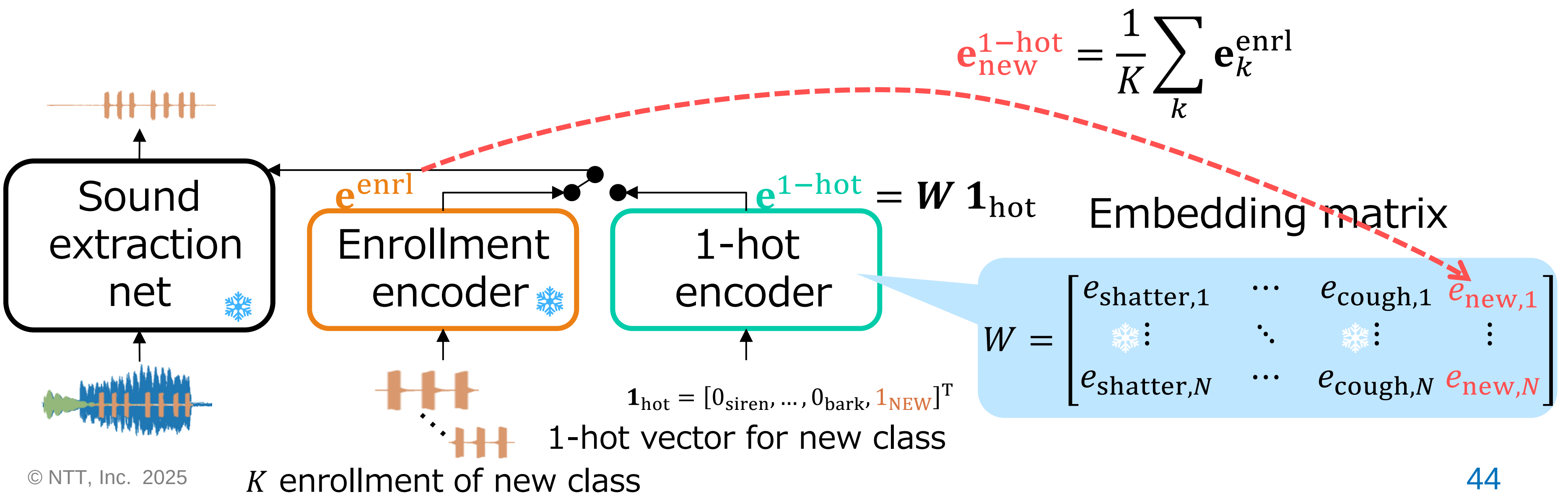
Goal:
- Learn new entries to the 1-hot embedding matrix for the new sound class
- Freeze all other parameters



Adaptation to new sound classes

We assume a few enrollment samples (e.g., K=5) from new sound class

1. Initialize 1-hot embedding of new sound classes with averaged embedding obtained from enrollment encoder

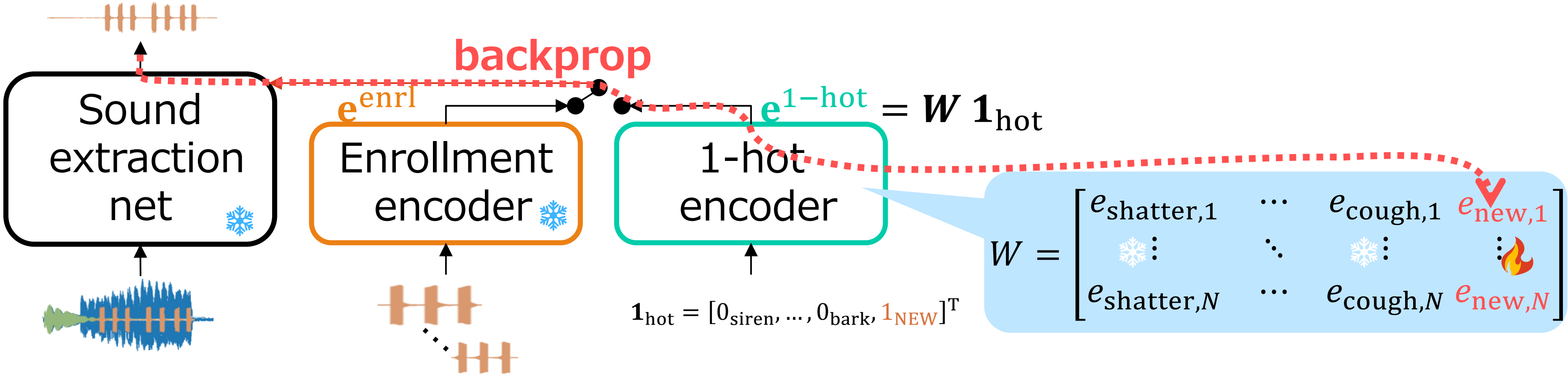


Adaptation to new sound classes

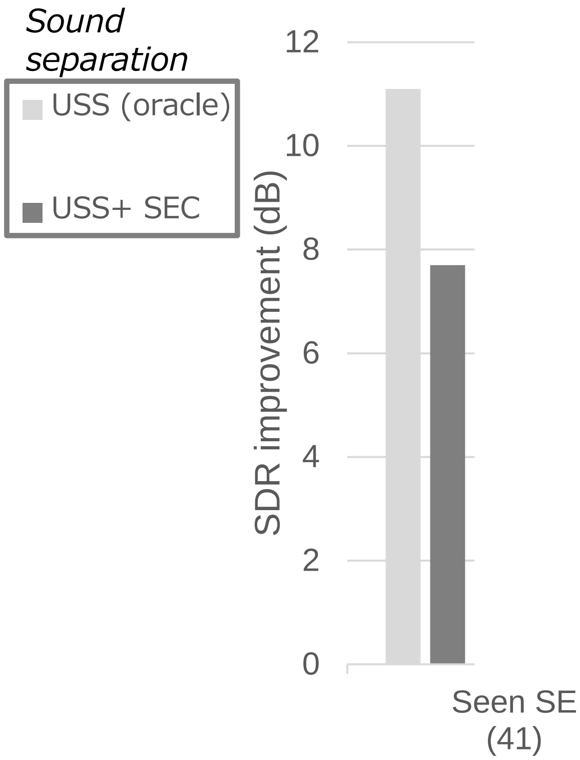
We assume a few enrollment samples (e.g., K=5) from new sound class

1. Initialize 1-hot embedding of new sound classes with averaged embedding obtained from enrollment encoder
2. Retrain only the 1-hot embedding by generating mixtures with few samples from the new sound classes

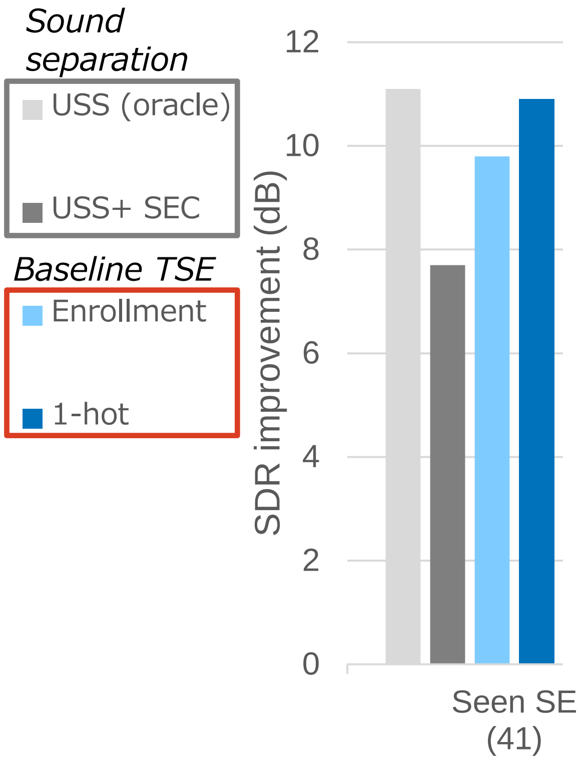
→ Allow continuous learning w/o catastrophic forgetting



Results on seen sound classes

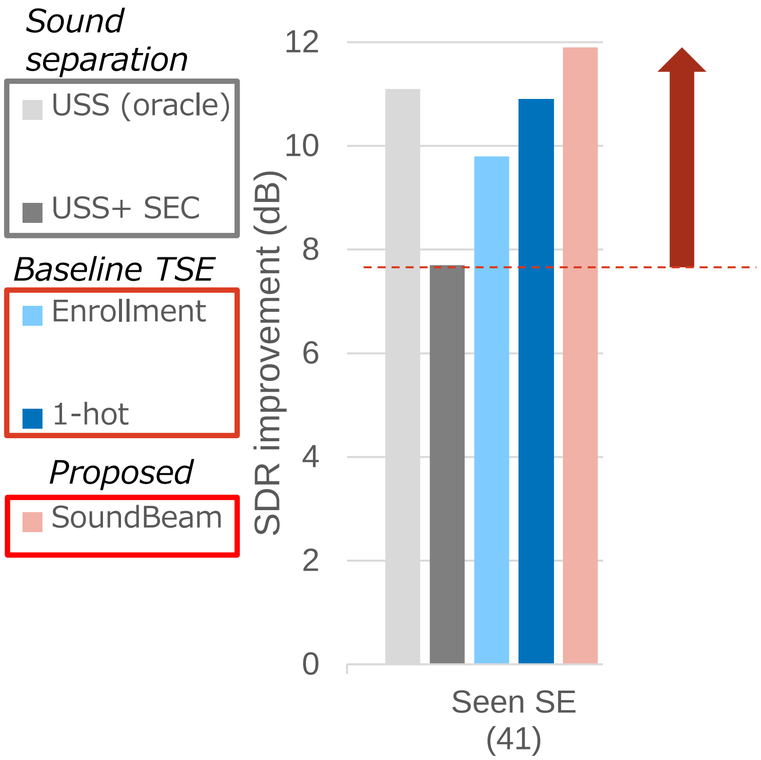


Results on seen sound classes



→ TSE outperforms cascade of sound separation with sound event classification (USS+SEC)

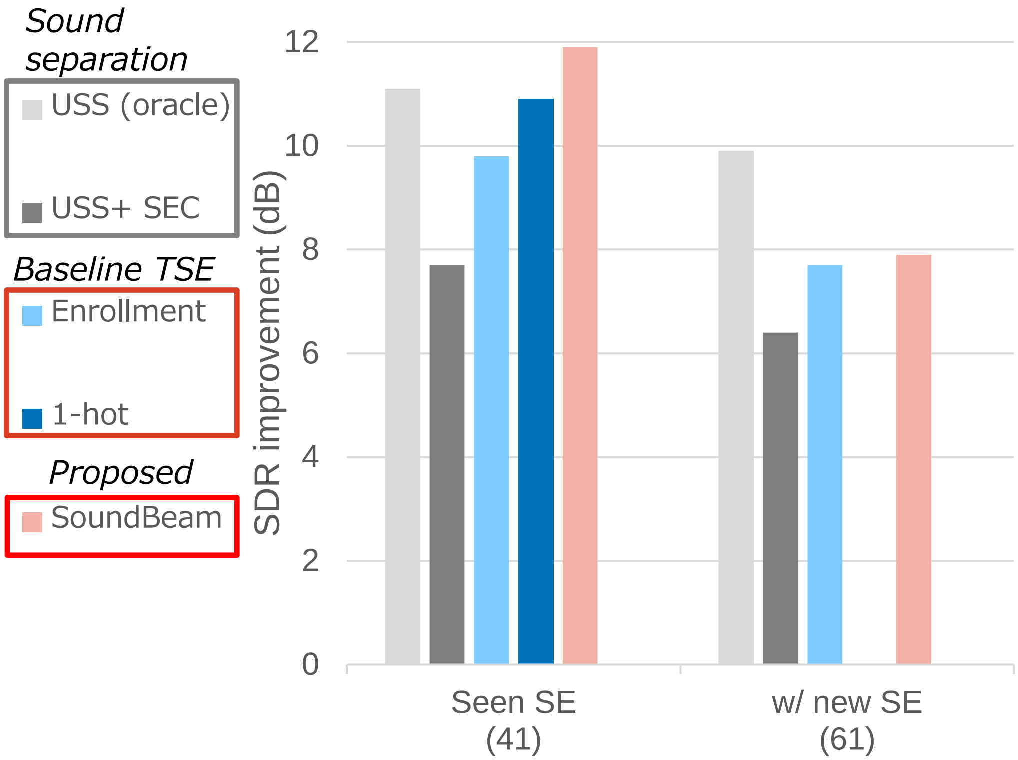
Results on seen sound classes



→ TSE outperforms cascade of sound separation with sound event classification (USS+SEC)

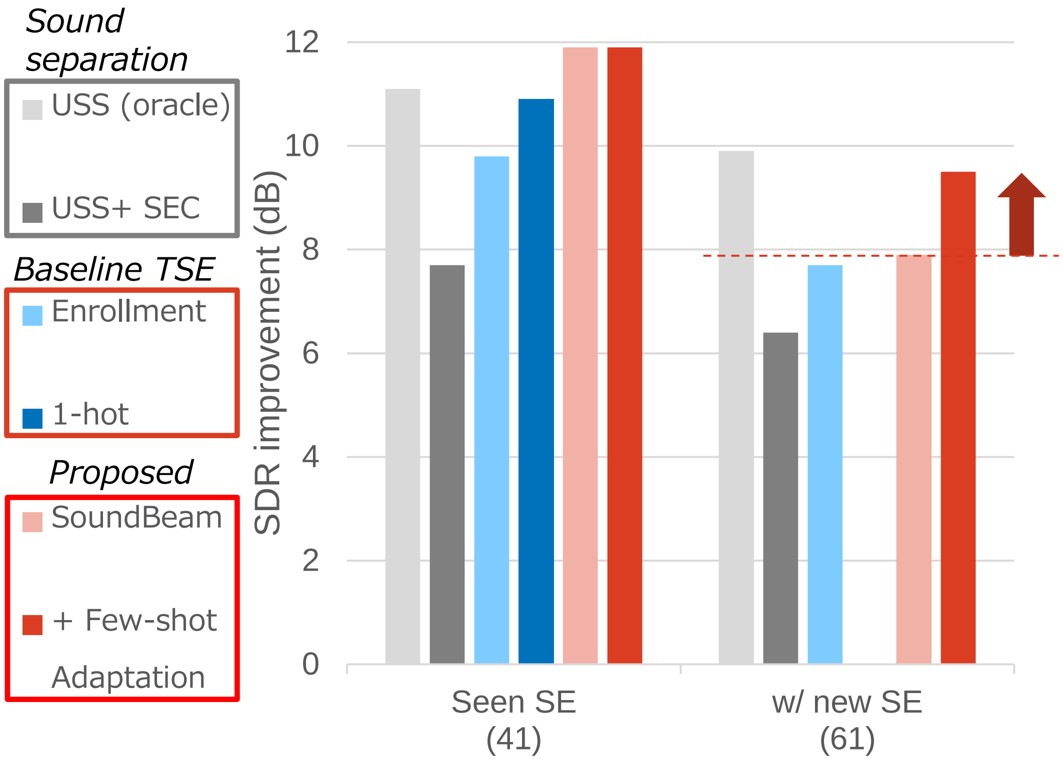
→ SoundBeam improves performance thanks to multi-task learning

Results for new sound classes



→ SoundBeam can handle new sound classes

Results for new sound classes



→ SoundBeam can handle new sound classes

→ Few-shot adaptation improves performance on new sound classes and maintain performance on seen classes

Wrap-up

- Can extend ideas of target speech extraction to arbitrary sounds
- Introduced a framework for continuous learning of sound classes

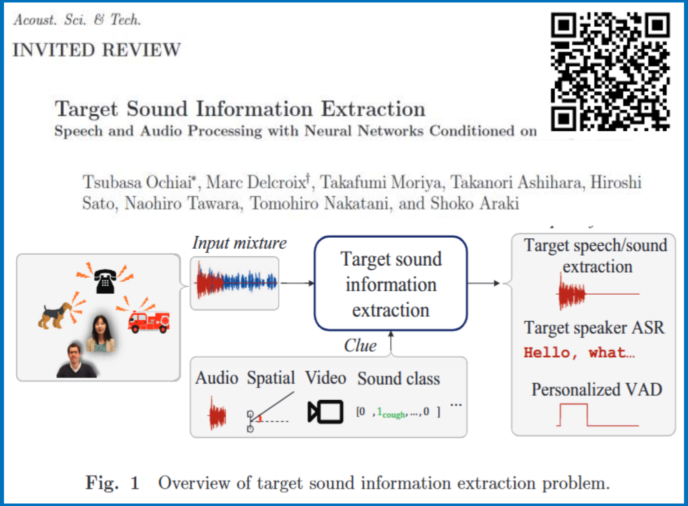
- Some remaining research directions
  - › TSE for smart hearable with lightweight/online processing
    - » B. Veluri, et al., “Real-time target sound extraction,” ICASSP, 2023.
    - » K. Wakayama et al., “Online Target Sound Extraction with Knowledge Distillation from Partially Non-Causal Teacher,” ICASSP, 2024.
  - › Improved performance for offline processing
    - » C. Hernandez-Olivan et al., "SoundBeam meets M2D: Target Sound Extraction with Audio Foundation Model," ICASSP 2025
  - › Other clues, e.g., sound description

NTT

Thank you!

Email: marc.delcroix@ieee.org

- K. Zmolikova, et al. *"Neural target speech extraction: An overview,"* IEEE Signal Processing Magazine 40.3 (2023): 8-29.
- M. Delcroix et al. , "*SoundBeam: Target Sound Extraction Conditioned on Sound-Class Labels and Enrollment Clues for Increased Performance and Continuous Learning,*" in IEEE/ACM TASLP, 2023.
- T. Ochiai et al. *"Target Sound Information Extraction: Speech and Audio Processing with Neural Networks Conditioned on Target Clues,"* AST 2025.


Acoust. Sci. & Tech.

INVITED REVIEW

Target Sound Information Extraction
Speech and Audio Processing with Neural Networks Conditioned on

Tsubasa Ochiai*, Marc Delcroix†, Takafumi Moriya, Takanori Ashihara, Hiroshi Sato, Naohiro Tawara, Tomohiro Nakatani, and Shoko Araki

Fig. 1 Overview of target sound information extraction problem.

Innovating a Sustainable Future for People and Planet

NTT