



Real-time TSE demonstration via SoundBeam with KD

Keigo Wakayama, Tomoko Kawase, Takafumi Moriya, Marc Delcroix, Hiroshi Sato,
 Tsubasa Ochiai, Masahiro Yasuda, Shoko Araki

NTT Corporation, Japan
 keigo.wakayama@ntt.com

Abstract

The objective of target sound extraction (TSE) is to extract sound sources of a specified class from mixed signals. Research into TSE has been actively conducted with the aim of applying it to immersive systems and auditory devices. We propose to demonstrate a real-time TSE system, which can isolate the signal from a desired sound class from sound mixtures recorded on the fly. This demonstration is based on the recently proposed causal SoundBeam model, which is trained using knowledge distillation (KD) from a non-causal TSE system. Experiments have demonstrated that SoundBeam with KD exhibits superior extraction accuracy compared to a state-of-the-art (SOTA) TSE, i.e., Waveformer. This paper explains the implementation of the proposed real-time TSE demonstration system. It is noteworthy that this demonstration will show for the first time at Interspeech, the ability to extract sound signals of a selected sound event (SE) class in real-time on a laptop.

Index Terms: Target sound extraction, SoundBeam, Causal model, Knowledge distillation, Real-time system, Demonstration

1. Introduction

Target sound extraction (TSE) is a process that can extract acoustic signals of a specific sound event (SE) class from a mixture of multiple sound signals. The potential applications of TSE are extensive, ranging from the domain of sound post-processing to the development of controllable hearing devices. TSE can be implemented with a deep neural network (DNN), which is conditioned on a clue that indicates the target SE class. Previous studies have investigated several clues, including SE class labels, audio queries from pre-recorded enrollment of the target sound classes, etc.

We have proposed a non-causal TSE approach called SoundBeam [1], which integrates enrollment-based TSE and class-based TSE and demonstrates superior performance due to the multi-task learning effect and generality to unseen classes. In a subsequent study, we have extended the non-causal SoundBeam to a causal SoundBeam using knowledge distillation (KD) from a partially non-causal teacher to mitigate performance degradation of the causal system [2]. This causal SoundBeam with KD has been shown to outperform a state-of-the-art (SOTA) TSE, i.e., Waveformer [3], which consists of dilated causal convolution and transformer decoders.

In this paper, we present the implementation of a real-time system for the causal SoundBeam with KD, and describe the demonstration we propose to present at the Interspeech 2025 “Show and Tell” session. As shown in Fig. 1, in this interactive demo, we record sound mixtures and process them in real-time, allowing participants to listen to the extracted signals from real

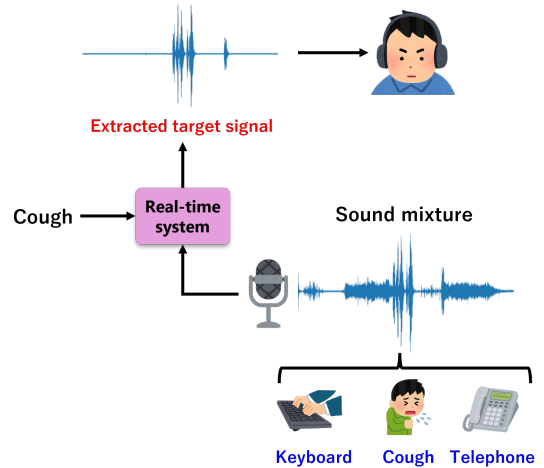


Figure 1: Demonstration scenario at “Show and Tell”.

mixtures. Target “speech” extraction demonstrations have been conducted at ICASSP 2019 [4] and ICASSP 2024 [5], but TSE demos have not been conducted at ICASSP and Interspeech. We are the first to demonstrate TSE from sound mixtures (with up to 61 target sound classes) in real-time at Interspeech.

2. Real-time SoundBeam with KD

SoundBeam [1] extracts sounds of a target SE class from a time-domain mixture signal, $\mathbf{y} \in \mathbb{R}^T$, containing several sounds that may overlap, where T is the signal duration. The target SE class is defined through the use of a 1-hot vector, $\mathbf{o}^{1\text{-hot}} \in \{0, 1\}^N$, or an enrollment audio sample, $\mathbf{a}^{\text{enroll}} \in \mathbb{R}^{T'}$. Here, N is the total number of classes. As shown in Fig. 2, the TSE process is described as:

$$\hat{\mathbf{x}}^s = \text{TSE}(\mathbf{y}, \mathbf{e}^s) \in \mathbb{R}^T, \quad (1)$$

where $\hat{\mathbf{x}}^s$ is an estimated target signal, $\text{TSE}(\cdot, \mathbf{e}^s)$ is a sound extraction model conditioned on an embedding vector, $\mathbf{e}^s \in \mathbb{R}^D$, and D is the feature dimension. We derive $\mathbf{e}^{1\text{-hot}}$ and $\mathbf{e}^{\text{enroll}}$ as \mathbf{e}^s using a 1-hot encoder as $\mathbf{e}^{1\text{-hot}} = f(\mathbf{o}^{1\text{-hot}})$, and an enrollment encoder as $\mathbf{e}^{\text{enroll}} = g(\mathbf{a}^{\text{enroll}})$. The $f(\cdot)$ and $g(\cdot)$ are the same as in Sec. IV. B of [1]. We train the system using the 1-hot and enrollment encoders alternatively. This improves performance due to the multi-task learning effect. Moreover, at inference, it allows using 1-hot for classes seen during training and enrollment for new “unseen” classes [1]. For simplicity, in the demonstration, we only perform TSE for seen classes.

All parameters are trained with an extraction loss consisting of the negative signal-to-distortion ratio [6], \mathcal{L}^{ext} .

