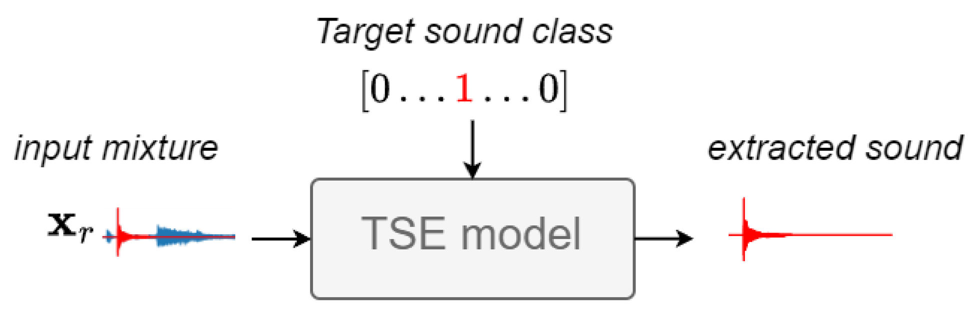
INTERAURAL TIME DIFFERENCE LOSS FOR BINAURAL TARGET SOUND EXTRACTION

Carlos Hernandez-Olivan, Marc Delcroix, Tsubasa Ochiai, Naohiro Tawara, Tomohiro Nakatani, Shoko Araki

NTT Corporation, Japan

Abstract

Introduction: Extract a source given a mixture.
Conventional TSE is single-channel:



Motivation

- Improve spatial metrics with new spatial loss formulation based on GCC-PHAT.

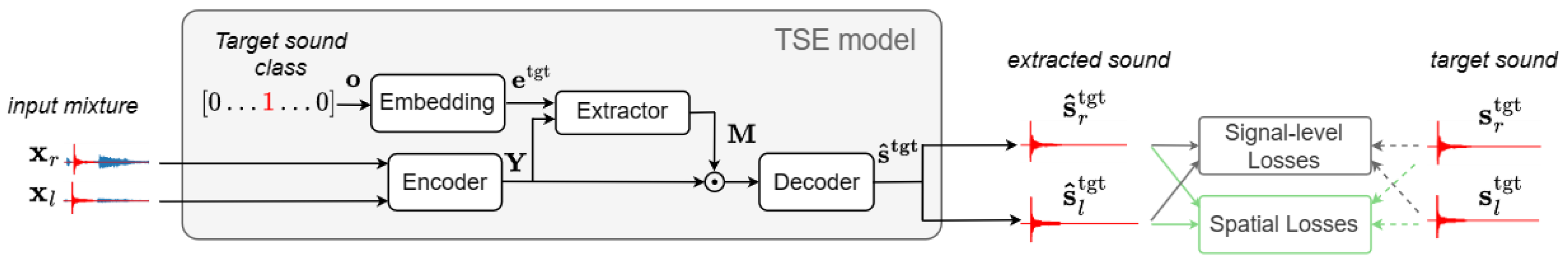
State-of-the-art

System	IPD loss	ILD loss	ITD loss
Semantic Hearing [1] (TSE)	✗	✗	✗
Tokala et al. [3] (speech enhancement)	✓	✓	✗
Ours (TSE)	✓	✓	✓

Contributions

- ITD loss that reduces both ITD and signal-level losses

Binaural Target Sound Extraction



$\hat{\mathbf{s}}^{\text{tgt}}$: extracted target sound signal
i : source index
m : microphone index
θ : network parameters

$\mathbf{x} = [\mathbf{x}_l, \mathbf{x}_r] \in \mathbb{R}^{T\times 2}$: binaural mixture
$\mathbf{s}^{\text{tgt}} = [\mathbf{s}_l^{\text{tgt}}, \mathbf{s}_r^{\text{tgt}}] \in \mathbb{R}^{T\times 2}$: binaural target
$\mathbf{o} = [0, \ldots, 1, \ldots 0]^\top$: one-hot vector
$r,\ l$: right and left channels

Observed signal

$$\mathbf{x}_m = \sum_{i=1}^{I} \mathbf{s}_{i,m}$$

T : number of samples
T' : number of frames
D : feature dimension

TSE Model

$$\hat{\mathbf{s}}^{\text{tgt}} = \text{TSE}(\mathbf{x}, \mathbf{o}; \theta) \qquad \mathbf{o} = [0, \ldots, 1, \ldots 0]^\top$$

$$\mathbf{Y} = \text{Encoder}(\mathbf{x}) \qquad \mathbf{M} = \text{Extractor}(\mathbf{Y}, \mathbf{e}^{\text{tgt}}) \qquad \hat{\mathbf{s}}^{\text{tgt}} = \text{Decoder}(\mathbf{M} \odot \mathbf{Y})$$

$$\mathbf{Y} \in \mathbb{R}^{T'\times D} \qquad \mathbf{e}^{\text{tgt}} \in \mathbb{R}^{D} \qquad \mathbf{M} \in \mathbb{R}^{T'\times D}$$

Training TSE

$$\hat{\theta} = \underset{\theta}{\text{argmin}}\, \mathcal{L}\left(\mathbf{s}^{\text{tgt}}, \hat{\mathbf{s}}^{\text{tgt}}(\theta)\right) \qquad \mathcal{L} = \alpha\mathcal{L}^{\text{signal}} + \beta\mathcal{L}^{\text{spatial}}$$

Signal-level and Spatial Losses

Signal-level Losses

$$\mathcal{L}^{\text{SNR}}(\mathbf{s}^{\text{tgt}}, \hat{\mathbf{s}}^{\text{tgt}}) = -\left(\frac{1}{2}\text{SNR}(\mathbf{s}_r^{\text{tgt}}, \hat{\mathbf{s}}_r^{\text{tgt}}) + \frac{1}{2}\text{SNR}(\mathbf{s}_l^{\text{tgt}}, \hat{\mathbf{s}}_l^{\text{tgt}})\right) \qquad \mathcal{L}^{\text{signal}} = 0.9\mathcal{L}^{\text{SNR}} + 0.1\mathcal{L}^{\text{SI-SNR}}$$

Spatial Losses

- Interaural level difference (ILD): difference in intensity between ears

$$\text{ILD} = 10\log_{10}\left(\frac{||\mathbf{s}_l||_2^2}{||\mathbf{s}_r||_2^2}\right) \qquad \boxed{\mathcal{L}^{\text{ILD}} = \left|\text{ILD}^{\text{tgt}} - \widehat{\text{ILD}}^{\text{tgt}}\right|}$$

- Interaural phase difference (IPD): difference in phase

$$\text{IPD}_{u,v} = \text{atan}\left(\frac{\text{Im}\left(S_{u,v,l}S^*_{u,v,r}\right)}{\text{Re}\left(S_{u,v,l}S^*_{u,v,r}\right)}\right) \qquad \boxed{\mathcal{L}^{\text{IPD}} = \frac{1}{UV}\sum_{u=1}^{U}\sum_{v=1}^{V}\left(\text{IPD}^{\text{tgt}}_{u,v} - \widehat{\text{IPD}}^{\text{tgt}}_{u,v}\right)^2}$$

- Interaural time difference (ITD): difference in arrival between ears

$$\mathbf{c} = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\mathbf{s}_l) \odot \mathcal{F}(\mathbf{s}_r)^*}{|\mathcal{F}(\mathbf{s}_l) \odot \mathcal{F}(\mathbf{s}_r)^*|}\right) \qquad \boxed{\text{ITD} = \underset{t\in[-\tau,\tau]}{\text{argmax}}\ c_t}$$

$$\mathcal{L}^{\text{ITD}} = \frac{1}{2\tau+1}\sum_{t=-\tau}^{\tau}\left(c_t^{\text{tgt}} - \hat{c}_t^{\text{tgt}}\right)^2$$

$\mathbf{c} = [c_{t=-T}, \ldots, c_{t=0}, \ldots, c_{t=T}] \in \mathbb{R}^{2T+1}$: vector of cross-correlation coefficients
c_t : cross-correlation coefficient between left and right channels
$\mathcal{F},\ \mathcal{F}^{-1}$: Fourier Transform (FT) and Inverse Fourier Transform (IFT)
$c_t^{\text{tgt}},\ \hat{c}_t^{\text{tgt}}$: cross-correlation between the reference and extracted signals
$\tau = 1\text{ms}$: scalar limiting the predicted delay in a range
$S_{u,v,r},\ S_{u,v,l} \in \mathbb{C}$: STFT of right and left channels $u,\ v$: indexes of time and frequency

Experimental Settings

Data: Semantic Hearing dataset [1]

- 20 sound classes from FSD50K, ESC-50, MUSDB18 and DISCO.
- HRTFS from CIPIC and real RIRs from 3 different corpora.
- 100K, 1K, 10K files for training, validation and test sets.
- Sampling frequency: 44.1KHz

Model Configuration

- Waveformer (E = D = 256) with lookahead.
- Encoder: 1-D convolution layer with stride of L = 32 samples and a kernel size of K = 3L.
- Extractor: DCC layers with kernel size 3 and dilation factors, and 2 MHA layers with 8 heads

Results

Table 1. Signal-level and spatial metrics for mixture, baseline TSE using only spatial loss and proposed systems using signal-level and spatial losses.

System		Signal-Level Metrics ↑ SI-SNR [dB]	↑ SNR [dB]	Spatial Metrics ↓ ΔILD [dB]	↓ ΔIPD [rad]	↓ ΔITD-GCC (ΔITD) [μs]	↓ FR [%]
(1)	Mixture	-0.74	-0.73	2.68	0.84	235.7 (263.0)	-
(2)	Baseline TSE w/ $\mathcal{L}^{\text{signal}}$	6.50	7.85	0.84	0.88	163.5 (86.3)	0.17
(3)	(2) + $\mathcal{L}^{\text{ILD}}$	6.72	8.10	**0.74**	0.83	168.5 (**74.8**)	**0.16**
(4)	(2) + $\mathcal{L}^{\text{IPD}}$	**6.76**	8.03	0.79	**0.49**	242.9 (80.1)	**0.16**
(5)	(2) + $\mathcal{L}^{\text{ITD}}$	6.74	**8.11**	0.78	0.84	**137.3** (79.0)	**0.16**

Conclusions

Conslusions

- The proposed ITD loss improves spatial metrics while not degrading signal-level losses.

Future Work

- Consider perceptual metrics and losses using high-pass filter for ILD and low-pass filter for ITD and IPD.

References

[1] Bandhav Veluri, Malek Itani, Justin Chan, Takuya Yoshioka, and Shyamnath Gollakota, “Semantic hearing: Programming acoustic scenes with binaural hearables,” in Proc. Symposium on User Interface Software and Technology (UIST). 2023, pp. 89:1–89:15, ACM.

[2] Bandhav Veluri, Justin Chan, Malek Itani, Tuochao Chen, Takuya Yoshioka, and Shyamnath Gollakota, “Real-time target sound extraction,” in Proc. ICASSP, 2023.

[3] Vikas Tokala, Eric Grinstein, Mike Brookes, Simon Doclo, Jesper Jensen, and Patrick A. Naylor, “Binaural speech enhancement using deep complex convolutional transformer networks,” in Proc. ICASSP. 2024, pp. 681–685, IEEE.