Sound Localization by Self-Supervised Time Delay Estimation (Extended Abstract)

Ziyang Chen, David F. Fouhey, and Andrew Owens

University of Michigan

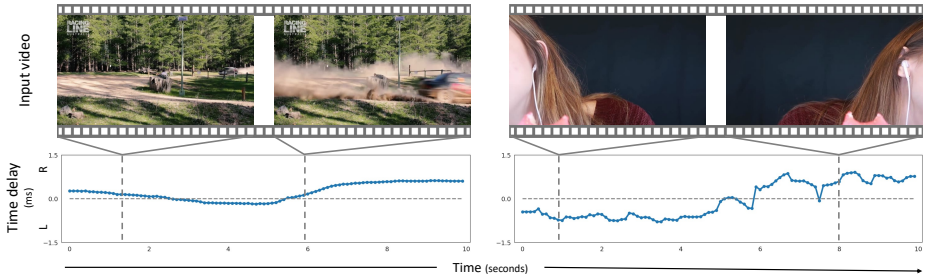

Fig. 1: Given a stereo audio recording, we estimate a sound's *interaural time delay*. Our model learns through self-supervision to find correspondences between the signals in each channel, from which the time delay can be estimated. We show time delay predictions for two scenes, along with their corresponding video frames (not used by the model). In both cases, the sound source changes its position in a scene, resulting in a corresponding change in time delay. The full paper can be found here: https://ificl.github.io/stereocrw.

1 Introduction

Sounds in the world arrive at one of our two ears slightly sooner than the other. This *interaural time delay*, which generally lasts only a few hundred microseconds, indicates a sound's direction and thus provides an important cue for multimodal perception. In humans, for example, time delays convey the positions of objects that move out of sight, and are integrated with visual cues when localizing events [16]. Visual information can also guide the sound localization process, allowing us to find a particular event of interest through binaural cues, while ignoring the others.

While high-quality stereo sound recordings are now abundant, existing methods [15,19,13,17,6] often struggle to localize sound sources within them and the difficulty in acquiring natural labeled data has limited their effectiveness. We propose to address these problems by learning time delay estimation from real, unlabeled recordings. Our approach, inspired by the *contrastive random walk* of Jabri et al. [14], learns audio embeddings that can be used to find *interaural correspondences* via cycle consistency: pairs of sounds from different stereo channels that correspond to the same underlying events. We show examples of time delay estimations for two real-world videos in Figure 1.

We also propose a model inspired by instance discrimination [7,20,11,3] that can perform a novel *visually-guided* time delay estimation task: localizing a speaker in a multi-speaker audio recording, given only their visual appearance. The resulting model is simple and can accurately localize speakers, without the need for explicitly separating sounds in the mixture.

Through experiments on simulated environments with metrically accurate ground truth, and on internet videos with directional judgments annotated by human listeners, we show that we can accurately estimate interaural time delays through self-supervised learning, using unlabeled stereo data and visual signals allow our models to localize specific speakers within mixtures.

2 Method

The goal of the time delay estimation problem is to determine how much sooner a sound reaches one microphone than another. Given the two channels of a stereo recording, $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$, represented as waveforms, and a function $h : \mathbb{R}^n \mapsto \mathbb{R}^{n \times d}$ that computes features for each temporal sample, a common solution is to choose a time delay τ that maximizes the generalized cross-correlation [15]:

$$R_{\mathbf{x}_1,\mathbf{x}_2}(\tau) = \mathbb{E}_t \left[\mathbf{h}_1(t) \cdot \mathbf{h}_2(t-\tau) \right], \tag{1}$$

where $\mathbf{h}_i = h(\mathbf{x}_i)$ are the features for $\mathbf{x}_i$, and $\mathbf{h}_i(t)$ is the d-dimensional feature embedding for time t.

Traditionally, the audio features, h, are defined using hand-crafted features, *e.g.*, Generalized Cross Correlation with Phase Transform (GCC-PHAT) [15]. We propose, instead, to learn h through self-supervision from unlabeled data.

2.1 Learning interaural correspondence

Our embeddings should provide *cycle consistent* matches: the process of matching features from $\mathbf{x}_1$ to those in $\mathbf{x}_2$ should yield the same correspondences as matching in the opposite direction, from $\mathbf{x}_2$ to $\mathbf{x}_1$.

We adapt the contrastive random walk model of Jabri et al. [14] to binaural audio. We create a graph that contains nodes for each of the temporal sample $\mathbf{x}_i(t)$ from both channels, with edges connecting the nodes that come from different channels. We then perform a random walk that transitions from nodes in $\mathbf{x}_1$ to those in $\mathbf{x}_2$, then back to $\mathbf{x}_1$, with transition probabilities that are defined by dot products between embedding vectors:

$$A_{ij}(s,t) = \frac{\exp(\mathbf{h}_i(s) \cdot \mathbf{h}_j(t)/c)}{\sum_{k=1}^{n} \exp(\mathbf{h}_i(s) \cdot \mathbf{h}_j(k)/c)}, \tag{2}$$

where $A_{ij}(s,t)$ is the probability of transitioning from sample s in $\mathbf{x}_i$ to sample t in $\mathbf{x}_j$, and a temperature constant c. The features $\mathbf{h}_i = h(\mathbf{x}_i; \theta)$ are parameterized with network weights θ and are represented using a ResNet [12]. We maximize the log return probability of a walk that moves between nodes in the two channels:

$$\mathcal{L}_{\texttt{crw}} = -\frac{1}{n} \operatorname{tr}(\log(A_{12}A_{21})). \tag{3}$$

2.2 Visually-guided time delay estimation

We also apply our model to the novel problem of estimating the time delay for a single sound within a mixture using visual information. Given a sound mixture containing multiple simultaneous speakers, we estimate the time delay for one object, given a visual representation of its appearance.

We adapt the instance discrimination [20] variation of our model, with a training procedure that resembles the "mix-and-separate" [21] paradigm. Given two audio tracks $\mathbf{u}$ and $\mathbf{v}$, we create a synthetic binaural sound mixture $\mathbf{x}_1 = \mathbf{u}+\mathbf{v}$ and $\mathbf{x}_2 = \text{shift}(\mathbf{u}, \tau_u) + \text{shift}(\mathbf{v}, \tau_v)$ for randomly sampled values τ_u and τ_v, where $\text{shift}(\mathbf{x}, \tau)$ shifts $\mathbf{x}$ by τ. The model is also provided with I_u, an image depicting $\mathbf{u}$. We learn audio-visual features by minimizing:

$$\mathcal{L}_{\texttt{av}} = -\log \frac{\exp(\mathbf{g}_1(t) \cdot \mathbf{g}_2(t-\tau_u)/c)}{\sum_{k=1}^{n} \exp(\mathbf{g}_1(t) \cdot \mathbf{g}_2(k)/c)}, \tag{4}$$

over all timesteps t, where $\mathbf{g}_i = g(\mathbf{x}_i, I_u)$ are the learned audio-visual features for channel $\mathbf{x}_i$.

2.3 Estimating delays from features

After learning our representation h or g, we can use it to estimate the time delay by maximizing $R_{\mathbf{x}_1,\mathbf{x}_2}$ (Eq. 1). Each embedding votes on a value for τ which can be performed by the nearest neighbor search, or by treating the learned similarities as probabilities (Eq. 2) and taking the expectation, *i.e.*, $\frac{1}{n}\sum_{s,\tau} \tau A_{12}(s,\tau)$ [2]. We then estimate final time delay from these votes, either by taking the mean or by using a RANSAC-like [8] mode estimation method.

3 Experiments

We evaluate our methods using both simulated audio with accurate time delays, and real-world binaural audio with quantized sound direction categories.

3.1 Evaluation of Interaural Correspondence

We train our audio-based model, **StereoCRW**, on datasets of stereo sound: FAIR-Play [9] and FreeMusic-Archive (FMA) [5] and compare it with GCC-PHAT [15], and the *supervised* method Salvati et al. [1].

Evaluation with simulated data. We first evaluate our models on simulated sounds with SNR = 10 and $\text{RT}_{60} = 0.5$s, a condition with moderate amounts of noise and reverberation. To measure prediction accuracy, we use mean absolute error (MAE) and root mean square error (RMSE) in milliseconds (ms). For all the methods, we provide 1024 (0.064s)

Table 1: **Delay estimation on simulated data.** We use SNR=10 and RT_{60}=0.5s. FAIR is FAIR-Play [9], FMA is FreeMusic-Archive [5]. *Vox-Sim* is the simulator [18]] with VoxCeleb2 [4] clips and *FMA-Sim* is the simulator with FMA clips. *Sup* refers to supervision, and *Aug* refers to augmentation.

Model	Variation	Data	Sup	Aug	MAE	RMSE
Salvati et al. [1]	Mean	Vox-Sim	✓		**0.126**	**0.254**
	Mean	Vox-Sim	✓	✓	0.169	0.294
	Mean	FMA-Sim	✓		0.135	0.256
	Mean	FMA-Sim	✓	✓	0.146	0.267
GCC-PHAT [15]	Mode	–			0.179	0.396
	Mean	–			**0.160**	**0.318**
Ours	Random	–			0.448	0.505
	StereoCRW	FAIR			0.241	0.364
	StereoCRW	FAIR		✓	0.174	0.322
	StereoCRW	FMA			0.434	0.654
	StereoCRW	FMA		✓	**0.133**	**0.259**

audio samples (at 16Khz) as input and perform 128 time delay prediction votes. As shown in Tab. 1, the StereoCRW model substantially outperforms GCC-PHAT when it is trained on FreeMusic-Archive, obtaining performance comparable with supervised models trained on synthetic data.

Evaluation with in-the-wild audio recordings. We then evaluate how well our method can localize sound directions in challenging real-world scenes, using audio collected from the internet where ground truth directions are annotated by human listeners. As shown in Tab. 2, our proposed approach substantially outperform GCC-PHAT and shows comparable results to a state-of-the-art *supervised* method, Salvati et al.

Table 2: **In-the-wild evaluation.** We evaluate our models' ability of localizing sounding objects on **in-the-wild** test cases.

Model	Variation	Aug.	Dataset	Acc (%) ↑
Salvati et al. [1]	Mean		Vox-Sim	87.0
	Mean	✓	Vox-Sim	87.3
	Mean		FMA-Sim	87.9
	Mean	✓	FMA-Sim	**89.1**
Chance	–		–	50.0
IID	–		–	75.5
GCC-PHAT [15]	Mean		–	81.3
Ours	Random	–	–	72.5
	StereoCRW		FMA	82.5
	StereoCRW	✓	FMA	**88.7**

3.2 Visually-guided Time Delay Estimation

We train our audio-visual model on VoxCeleb2 [4] with paired face images and mono audio. We evaluated methods on simulated data using two metrics: RMSE, and the percentage of predictions with less than 0.1 ms (1.6 samples) of error (Err$\leq$ 0.1). We feed models 1.0s or 2.55s audio, resulting in 512 votes. We compare our audio-visual approach with audio-only methods, a (oracle) baseline which selects one of the two speakers' ground-truth time delay at random, and a two-stage method that first separates the speaker's voice for each channel using VisualVoice [10], then applies audio-based time delay estimation methods to the separated sounds. We show results in Tab. 3. Our audio-visual model substantially outperforms the audio-only baselines and performs comparably to Sep+GCC-PHAT in regression metrics.

Table 3: **Visual-guided time delay estimation on simulated data.** We evaluate our models' ability of predicting ITD signals from mixtures with visual aids.

Audio duration	0.96s		2.55s	
Model	RMSE	Err$\leq$ 0.1 ↑	RMSE	Err$\leq$ 0.1 ↑
GCC-PHAT [15]	0.503	56.6	0.504	56.9
Salvati et al. [1]	0.490	52.5	0.483	50.1
Random Oracle	0.502	56.9	0.502	56.9
Ours - Random	0.493	10.0	0.503	9.76
Ours - StereoCRW	0.493	56.8	0.488	55.7
Ours - AV	**0.304**	**72.5**	**0.295**	**76.1**
Sep [10] + GCC	0.361	77.6	0.323	82.2
Sep [10] + StereoCRW	**0.309**	**82.8**	**0.281**	**85.5**

Real-world visually-guided localization. We also perform experiments on a self-recorded video (Fig. 2). Two speakers talk concurrently while moving off-screen. Our model localizes each speaker in the mixture with a cropped image of their face. We show the mean and standard deviation of delay predictions in 2.0s windows.

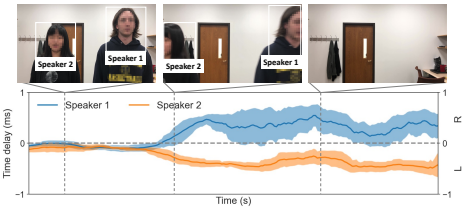

Fig. 2: **Visually-guided localization for a real-world scene**.

References

1. Time delay estimation for speaker localization using cnn-based parametrized gcc-phat features 3, 4
2. Bian, Z., Jabri, A., Efros, A.A., Owens, A.: Learning pixel trajectories with multi-scale contrastive random walks. arXiv (2022) 3
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020) 2
4. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018) 3, 4
5. Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X.: Fma: A dataset for music analysis. arXiv preprint arXiv:1612.01840 (2016) 3
6. Diaz-Guerra, D., Miguel, A., Beltran, J.R.: Robust sound source tracking using srp-phat and 3d convolutional neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 300–311 (2020) 1
7. Dosovitskiy, A., Fischer, P., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE transactions on pattern analysis and machine intelligence **38**(9), 1734–1747 (2015) 2
8. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM (1981) 3
9. Gao, R., Grauman, K.: 2.5 d visual sound. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 324–333 (2019) 3
10. Gao, R., Grauman, K.: Visualvoice: Audio-visual speech separation with cross-modal consistency. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15490–15500. IEEE (2021) 4
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. arXiv preprint arXiv:1911.05722 (2019) 2
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition (CVPR) (2016) 2
13. Houegnigan, L., Safari, P., Nadeu, C., van der Schhaar, M., Solé, M., Andre, M.: Neural networks for high performance time delay estimation and acoustic source localization. In: Proceedings of the Second International Conference on Computer Science, Information Technology and Applications. pp. 137–146 (2017) 1
14. Jabri, A., Owens, A., Efros, A.A.: Space-time correspondence as a contrastive random walk. arXiv (2020) 1, 2
15. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. IEEE transactions on acoustics, speech, and signal processing **24**(4), 320–327 (1976) 1, 2, 3, 4
16. Kumpik, D.P., Campbell, C., Schnupp, J.W., King, A.J.: Re-weighting of sound localization cues by audiovisual training. Frontiers in Neuroscience **13**, 1164 (2019) 1
17. Pertilä, P., Parviainen, M.: Time difference of arrival estimation of speech signals using deep neural networks with integrated time-frequency masking. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 436–440. IEEE (2019) 1
18. Scheibler, R., Bezzam, E., Dokmanić, I.: Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 351–355. IEEE (2018) 3

19. Schmidt, R.: Multiple emitter location and signal parameter estimation. IEEE transactions on antennas and propagation **34**(3), 276–280 (1986) 1
20. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3733–3742 (2018) 2, 3
21. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European conference on computer vision (ECCV). pp. 570–586 (2018) 3