The Cone of Silence
Speech Separation by Localization

Teerapat Jenrungrot*, Vivek Jayaram*, Steven M. Seitz, Ira Kemelmacher-Shlizerman
University of Washington, Seattle

UW REALITY LAB

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

NEURAL INFORMATION PROCESSING SYSTEMS

Project page: https://grail.cs.washington.edu/projects/cone-of-silence/

Introduction

Given a multi-microphone recording of an unknown number of speakers talking concurrently, we simultaneously localize the sources and separate the individual speakers. At the core of our method is a deep network, in the waveform domain, which isolates sources within an angular region $\theta \pm \frac{w}{2}$, given an angle of interest θ and angular window size $\mathcal{W}$. By exponentially decreasing $\mathcal{W}$, we can perform a binary search to localize and separate all sources in logarithmic time. Our algorithm allows for an arbitrary number of potentially moving speakers at test time, including more speakers than seen during training. Experiments demonstrate state-of-the-art performance for both source separation and source localization, particularly in high levels of background noise.

Method

To specify an angle of interest θ to the network, we shift the multi-channel audio signal based on the specified location. The shift amount for each channel can be computed based on the distance between source and microphone, speed of sound and sampling rate, using the following equation.

$$T_{\texttt{delay}}(p_\theta, \texttt{mic}_i) = \left\lfloor \frac{d(p_\theta, \texttt{mic}_i)}{c} \cdot sr \right\rfloor$$

We specify an angular window size $\mathcal{W}$ by globally conditioning the network using the one-hot vector corresponding to the window size. Figures 1 and 2 show our network architecture, adapted from [1].

Given an angular separation network, we separate and localize all voices in a single scene by exponentially decreasing the angular window size $\mathcal{W}$. Initially, our method runs the separation network on each quadrant and respectively subdivides the search space until we have fine-grained localization resolution. Figure 3 demonstrates our separation and localization method.

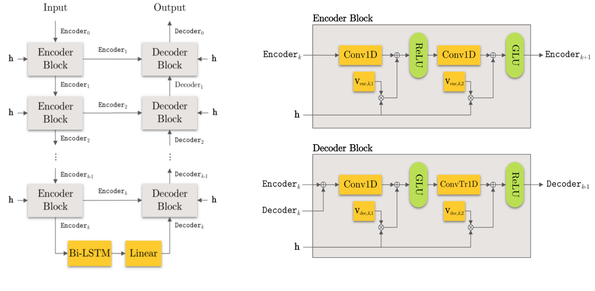

Figure 1 (left): Our network architecture.

Figure 2 (right): The encoder and decoder blocks. In both figures, h refers to the global conditioning variable corresponding to an angular window size w.

Input Scenario

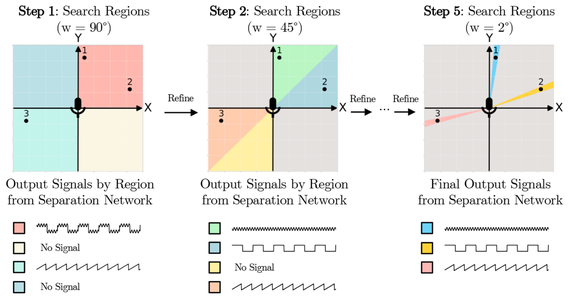

Figure 3: Overview of *Separation by Localization* running binary search on an example scenario with 3 sources. Each panel shows the spatial layout of the scene with the microphone array located at the center. During Step 1, the algorithm performs separation on candidate regions of 90°. The quadrants with no sound get suppressed and disregarded. The algorithm continues doing separation on smaller partitions of candidate regions until reaching the final step where the angular window size is 2°.

Experimental Results

Source Separation

Our method shows state-of-the-art separation performance in a simple scenario with 2 voices and 1 background. Our method strongly outperforms the best possible results obtainable with spectrogram masking and is slightly better than recent deep-learning baselines [2,3] operating on the waveform domain. Additionally, our method can accept explicitly known source location, slightly improving the separation performance (denoted as Ours - Oracle Location)

Method	**SI-SDRi** (dB)
Waveform-based	
Conv-TasNet [18]	15.526
TAC [40]	15.121
Ours - Binary Search	**17.059**
Ours - Oracle Location	17.636
Spectrogram-based	
Oracle IBM [64, 65]	13.359
Oracle IRM [64, 66]	4.193
Oracle MWF [64, 67]	8.405

Table 1 (left): Separation performance. Larger SI-SDRi is better. The SI-SDRi is computed by finding the median of SI-SDR increases evaluated on our synthetic dataset.

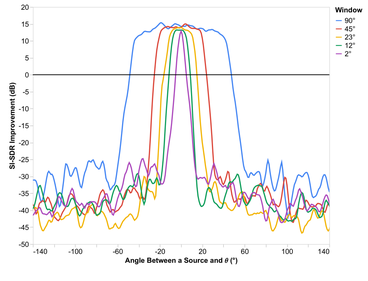

Figure 4 (right): Evidence that the network amplifies voices between $\theta \pm \frac{w}{2}$ and suppresses all others.

Source Localization

Our method shows state-of-the-art performance in the simple scenario with 2 voices, but some baselines show similar performance to ours. However, when background noise is introduced, the gap between our method and the baselines increase greatly. Learning-based baseline, MLP-GCC, [4] struggles to distinguish a voice location from background noise.

Method	**Median Angular Error**	
	2 Voices	2 Voices + BG
Learning-free		
MUSIC [33]	82.5°	36.8°
SRP-PHAT [32]	6.2°	46.4°
CSSM [34]	30.1°	36.3°
WAVES [35]	16.4°	32.1°
FRIDA [37]	6.9°	18.5°
TOPS [36]	2.4°	11.5°
Learning-based		
MLP-GCC [38]	1.0°	41.5°
Ours	**2.1°**	**3.7°**

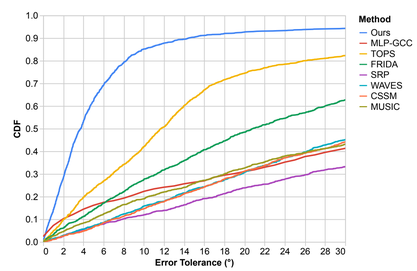

Table 2 (left): Localization performance.

Figure 5 (right): Error tolerance curve on mixtures with 2 voices and 1 background.

Generalization to high number of speakers

Number of Speakers N	2	3	4	5	6	7	8
SI-SDRi (dB)	13.9	13.2	12.2	10.8	9.1	7.2	6.3
Median Angular Error	2.0°	2.3°	2.7°	3.5°	4.4°	5.2°	6.3°
Precision	0.947	0.936	0.897	0.912	0.932	0.936	0.966
Recall	0.979	0.972	0.915	0.898	0.859	0.825	0.785

Table 3: Generalization to arbitrary many speakers. We report the separation and localization performance as the number of speaker varies. For this experiment, we trained a network using scenes with only up to 4 speakers.

Real-world Scenarios

In the supplementary videos, we explore a variety of real-world scenarios. These include multiple people talking concurrently and multiple people talking while moving.

Acknowledgements

The authors thank the labmates from UW GRAIL lab. This work was supported by the UW Reality Lab, Facebook, Google, Futurewei, and Amazon.

References

[1] Alexandre Défossez et al. (2019). "Demucs: Deep extractor for music sources with extra unlabeled data remixed." In arXiv preprint arXiv:1909.01174, 2019.

[2] Yi Luo et al. (2019). "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation." In: IEEE Transactions on Audio, Speech, and Language Processing, pp.1256-1266, 2019.

[3] Yi Luo et al. (2020). "End-to-end microphone permutation and number invariant multi-channel speech separation." In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.6394-6398, 2020.

[4] Weipeng He et al. (2018). "Deep neural networks for multiple speaker detection and localization." In: IEEE International Conference on Robotics and Automation (ICRA), pp.74-79, 2018.