Spatial-Aware Multi-Task Learning Based Speech Separation

Wei Sun*, Mei Wang*, Lili Qiu++

The University of Texas at Austin* Microsoft Research Asia++

Abstract—Online meetings have become an indispensable part of our lives. However, background noise from other family members, roommates, and office mates not only degrades the voice quality but also raises serious privacy issues. In this paper, we develop a novel system, called Spatial Aware Multi-task learning-based Separation (SAMS), to extract audio signals from the target user during teleconferencing. Our solution consists of two novel components: (i) generating fine-grained location embeddings from the user’s voice and inaudible tracking sound, which contains the user’s position and rich multipath information, (ii) developing a source separation neural network using multi-task learning to jointly optimize source separation and location.

Index Terms—Mobile Application, Acoustic Sensing

I. INTRODUCTION

Online meetings play an indispensable role in our daily life. Many kids depend on it for education, adults rely on it for work, and friends count on it for socialization. Over 70% US households have two or more people and the average household size across the world is 4. While one is participating in an online meeting or taking an online class, other house members may generate sound. Unlike videos, voice signals can travel across rooms and result in significant interference unintentionally. This both degrades the audio quality and raises serious privacy issues [1].

Existing work: There has been significant work on signal source separation. Earlier works use signal processing, such as principal component analysis (PCA) and independent component analysis (ICA). More recent works use machine learning (ML) to further improve separation accuracy. Videos can also be used to improve source separation [2]–[4] since the camera captures mouth position and movement. However, video requires good lighting conditions, has a limited field of view, and raises significant privacy concerns. [5] also reports turning off a camera during teleconferencing reduces the environmental footprint of a meeting by 96%.

Despite considerable work, existing works primarily focus on using raw audio samples for source separation. User location can have a significant impact on source separation but has not been explicitly considered until recently. Some recent works use the ground truth location for source separation and report significant benefits [6], [7]. The existing video/audio work also assumes the mouth location is accurate, hence it cannot be directly applied to audio-only solutions since acoustic tracking has a larger error. CoS [8] uses a binary search for the azimuth direction, which may fail due to multipath cancellation in a wider beam.

Let’s consider a user using a computer to join an online meeting while other people are talking in the background. From Figure 1, we observe that the user is moving while speaking. Although the absolute distance change is small in Figure 1(b), it can result in a large fluctuation in the Angle of Arrival (AoA) as shown in Figure 1(a) because the user is close to the computer during the meeting and even a small movement may cause a large AoA change. Our measurements in Figure 1 show that there is a frequent head movement that yields over a 20-degree change in the AoA and a 20 cm change in the distance when a user speaks spontaneously. Therefore, it is useful to track the user’s location and use it for source separation. Some users prefer to perform other tasks during the teleconference with a pair of Bluetooth earphones. This is a less common case and out of the scope of this paper.

To fully exploit the spatial information, we develop a system SAMS that automatically removes the interference by explicitly estimating the user location and multipath profiles and using the estimated spatial embeddings to enhance source separation accuracy as shown in Figure 2. The smart speaker receives both audible and inaudible reflections. SAMS later analyzes the spatial information and semantic knowledge from the acoustic signal and uses multi-task learning to jointly learn the AoA and speech separation.

While there has been existing work on localizing a user using inaudible signals [9], [10] or audible signals [11], [12], our work is the first that leverages both audible and inaudible signals to achieve high localization accuracy. More specifically, we extract masks in the Time-Frequency (TF) domain from audible signals and use the mask to select TF bins dominated by the target user’s voice to improve the localization accuracy of the voice signal under interference. Meanwhile, the device emits inaudible chirps to track position. We feed location profiles from both audible and inaudible signals to generate spatial embeddings, which will serve as the input to source separation. These spatial embeddings contain not only the user’s location but also rich multipath information and play an important role in source separation.

We leverage the user’s location to improve the source separation by developing a novel multi-task learning framework to jointly learn the source separation and location. Instead of
Multiple-channel based separation: Animals have developed multiple ears through millions of years of evolution. Similarly, more receiver channels can significantly improve separation performance both in theory and in practice. Beamforming algorithms [19] leverage spatial information to strengthen the signal in target directions and null the interference from unwanted directions. [6] proposes to iteratively run classic beamforming and separation to guide the network to focus in the appropriate direction. Recent work [20] develops end-to-end learning of complex covariance matrix to predict spatial filter. These works require the source position as an additional input for source separation. Concurrent work [21] develops binaural separation systems on earphones. They implicitly use the head-related transfer function (HRTF) to separate sources. This targets earphones while our work targets regular speakers and microphones, so it is complementary.

Leveraging context information: User positions contain important cues for interference cancellation. CoS [8] integrates a binary search of the azimuth direction with separation. A well-known issue of binary search is that multipath signals may be canceled out in a wider cone that includes the real AoA, which will prevent it from zooming in the right direction and missing the real AoA. [22] develops a wearable directional hearing system but it still requires direction as the input. Ultrasound is also proposed for source separation in [4], [23], but the hardware is not widely available. UltraSE [24] and WaveVoice [25] emit the inaudible sound to the mouth to measure the Doppler shift or the modulated phase and amplitude information of lip movement within 0.2 m. WaveVoice [25] shows significant performance degradation when the distance to the phone is beyond the threshold. Our work targets computer users within 1 m away from the speaker/mic. It is hard to detect lip movements using inaudible acoustic signals in this range. Therefore, we focus on tracking user position using audible and inaudible signals. [26], [27] use mmWave to sense vocal vibration as a reference of target user position using audible and inaudible signals. [26], [27] use mmWave to sense vocal vibration as a reference of target user position using audible and inaudible signals. [26], [27] use mmWave to sense vocal vibration as a reference of target user position using audible and inaudible signals.
A natural aggregation of the MUSIC spectrum across different frequency bins is summing up MUSIC profiles from all TF bins and selecting the angle corresponding to the highest peak. However, not all TF bins contain the target user’s speech due to the sparsity of human speech over the frequency band [30]. Therefore, it is important to select the TF bins that contain strong Signal to Noise Ratios (SNR) from the target user. It is challenging to select the TF bins by just analyzing the power and phase. Thus, we propose to leverage semantic information

**Mask generation:** A number of approaches have been proposed to generate TF masks for speech enhancement. A few works estimate the amplitudes of audio spectrogram (e.g., real valued ideal binary mask (IBM) [31] and ideal ratio mask (IRM) [32]). [14], [30] develop DNN based approaches to generate amplitude and phase masks.

We use IBM to select appropriate TF bins for the target user from mixed noisy complex spectrogram. IBM is a method for speech separation based on deep neural network [31]. IBM is based on the sparsity of human speech (i.e., the number of non-silent TF bins from a speaker tends to be small). It determines a binary mask for each TF bin, where 1 means the target signal dominates interference and 0 otherwise. It decomposes downsampled audio signals into 2D TF bins. Then it extracts several features, such as autocorrelation of a filter response, autocorrelation of an envelop of filter response, and cross-channel correlation. Next, it performs clustering based on these extracted features, and tag each TF bin with either target dominant or interference dominant based on similarity with the clean target signal (spoken at a different time), which is also an input. We use the clean target signal, which is location independent and can be collected only once during user account creation. Since an online meeting requires a user to sign in, it is reasonable to assume the target user is known.

We train the IBM mask estimator using Deep Cluster [16], which is a general and robust method to estimate the mask. It takes a log power spectrogram (LPS) as the input to estimate an initial binary mask of the target user. We get the pre-trained model using the LibriMix [33] data along with our own testbed traces described in Section VII-A. The binary mask is a coarse estimate of effective TF bins, but it maintains phase information and is useful for selecting the TF bins for further analysis. We estimate binary masks for all microphones. To minimize interference, we select the TF bins for AoA estimation only when the masks from all microphone channels are 1s. In this way, we effectively remove the TF bins with large interference and noise.

**Applying a mask to MUSIC:** We apply the MUSIC algorithm to the TF bins with masks. Then we concatenate the MUSIC spectrum from all frequencies together. The output profile is represented as a 2D matrix of size $M_f \times N_a$ across different frequencies, where $M_f$ denotes the number of frequency bins and $N_a$ denotes the number of angles. $M_f$ is set to 103 frequency bins (equally spaced from 800Hz to 4KHz for human speech), and $N_a$ is set to 181 (spanning 0 degrees to 180 degrees with 1 degree apart) in our evaluation. This will...
be further combined with the output from inaudible signals for generating location embeddings.

V. LEVERAGING INAUDIBLE SENSING

Apart from audible band, we leverage the inaudible acoustic signal to improve the robustness of spatial representation.

A. Inaudible Side Channels

Estimating multipath profile solely based on speech has several limitations because of the audible band. First, audible signals may contain significant interference and ambient noise, which results in significant AoA errors. Figure 4 shows the AoA estimation from the audible target speech is fairly accurate without interference. However, when the interference is introduced to the audible band, the estimation deviates from the ground truth a lot. The deviation can be large and irregular due to unexpected speech from an unknown direction. To quantify the impact of interference, we apply MUSIC to the audible signals under different signal-to-interference and noise ratios (SINR). As shown in Figure 6(b), adding interference increases the AoA error significantly. The AoA error increases by 10.45° over no interference while SINR = -6 dB; the error increases by 13.53° when SINR = -12 dB. While using masks removes a significant amount of interference, the removal alone is insufficient to support accurate AoA estimation.

Besides, most energy in speech concentrates in low-frequency bands (e.g., below 2 kHz), which have large wavelengths and lead to low AoA resolution [11]. Furthermore, typically a relatively large time window (e.g., hundreds of ms) is used to analyze audible signals in order to ensure there is enough energy from the target speaker. This limits the update rate of multipath profile estimation. The temporal resolution of TF bins is 10 ms, but the AoA update rate cannot keep up with the TF bin update rate. We need more frequent multipath profiles encoded by the AoA estimation.

In comparison, inaudible signals are not affected by the interference and noise in the audible band. It can work as a side channel to detect and track the target speech. Meanwhile, inaudible signals have shorter wavelengths so that they can achieve higher AoA resolution. Finally, inaudible signals are modulated at the transmitter end. It can be designed as a small duration to improve the estimation rate. Hence, inaudible signals enable tracking at a much higher frequency (e.g., tens of ms), which is important to adapt more quickly to the changing user position.

However, in the typical usage scenarios where the speaker and microphones are on the desk, inaudible signals are mostly reflected by the user’s body instead of the mouth. Therefore, inaudible signals mainly track body movements. Besides, there are multiple reflections from the human body. It is challenging to distinguish which parts of the human body reflect the inaudible signal. Fortunately, body movement is highly correlated with mouth movement. As shown in Figure 4, the AoAs estimated using inaudible signals follow a similar moving trend to the ground truth even though its absolute AoA differs from the ground truth. Moreover, even though we may see reflections from different body parts, they tend to have a similar trend as the entire body moves together. The direction of the reflected inaudible signal is correlated with the direction of the speech. The other unrelated reflections from the environment can be filtered out by the spatial embedding, such as the static reflections from the wall or the dynamic reflections from moving objects. Figure 5 shows even though there are multiple reflections from the human body, these reflections share a similar moving pattern.

Thus inaudible signals can capture the motion and help correct inaccurate AoA estimation from the noisy audible bands. There may be multiple AoA candidates from inaudible reflection (e.g., due to reflection from multiple body parts), but they have similar movement trends, which are useful for tracking and source separation. We develop a neural network to automatically exploit the features extracted from both audible and inaudible signals.

B. Generating Spatial Embeddings

Feature extraction with 2D MUSIC: We let a speaker on a computer transmit periodic FMCW chirps, whose frequency increases linearly from $f_{min}$ to $f_{max}$ during each period $T$. This signal enables distance estimation as well as AoA estimation with a microphone array. We use 2D MUSIC [34] to generate a distance and AoA profile as the inaudible representation. Each MUSIC profile can be considered as an image, and a sequence of MUSIC profiles can thus be treated as a video sequence. We apply 3D convolution [35] with $5 \times 7 \times 7$ kernels and ResNet-18 [36] to MUSIC profiles generated from audible and inaudible signals to effectively learn spatio-temporal...
features. The temporal window of 3D kernels helps filter out noisy patterns based on neighbor frames. It is followed by batch normalization and ReLu activation. Then output features are fed to ResNet18 to encode profile embeddings with 512 dimensions. The concatenated embeddings from audible and inaudible profiles will be directed to the source separation.

A sequence of embeddings are extracted from the MUSIC profiles. We feed them into both LSTM and source separation networks. LSTM takes these embeddings to estimate AoAs because they include important spatial information from audible and inaudible signals over the recent time window. The 3D convolutional layers use temporal information in small time windows at an early stage, while the LSTM can leverage a much longer sequence. Moreover, audible embedding and inaudible embedding complement each other due to the correlation between mouth movement and body movement. The memory unit in LSTM helps track long-term movement. Our objective is to minimize the L1 loss between the estimated AoA (AoA) and ground truth AoA (AoA), denoted as $L_{AoA} = ||\hat{AoA} - AoA||_1$. We use LSTM to learn the temporal features from the audible and inaudible location embeddings. Each cell is followed by a linear layer with 128 input nodes and 1 output node, which is the estimated AoA. There is 1 hidden linear layer with 64 nodes and ReLU.

VI. MULTI-TASK LEARNING FRAMEWORK

Considering the strong inter-dependency between tracking and source separation, we apply a multi-task learning framework to jointly estimate the location and separate the source. In our context, predicting the AoA is one task while separating the target speech is another task. In the context of this paper, we jointly learn AoA estimation and speech separation at the same time. The key observation is that AoA estimation and speech separation can benefit from each other. Both tasks require shared spatial representation as the key feature.

AoA estimation benefits separation. Beamforming can be used for source separation. It combines signals from multiple antennas to strengthen desired signals and null undesired signals in the other directions. We illustrate the impact of AoA estimation on beamforming performance using MVDR as an example. MVDR is an adaptive algorithm that minimizes interference and noise while preserving the desired signal from a given direction. Its beamforming weight is $w(f, AoA) = \frac{\Phi^{-1}_{\hat{AoA}}(f)w(f, AoA)}{\Sigma_{t=0}^{T}\Phi_{\hat{f}}^{-1}(f)w(f, AoA)}$, where $\Phi^{-1}_{\hat{AoA}}$ is the covariance matrix of the target speech at the frequency band $f$, and $w(f, AoA)$ is the steering vector of the target speech at the frequency.

The goal of source separation is to minimize the difference between beamformed speech $\hat{s}$ and target speech $s$. The formulation VI reveals that the AoA is the key factor to construct the steering vector and determine the beamforming weights. Figure 6(a) shows that a more accurate AoA estimation yields better beamforming and source separation as well.

Separation helps AoA estimation. We observe that a clean reference signal can improve the AoA estimation in various algorithms. Figure 6(b) shows the AoA error under different signal-to-interference-and-noise ratios using the MUSIC algorithm. The clean signal contains accurate phase difference across multiple channels and generates an accurate covariance matrix for the MUSIC algorithm to estimate the AoA. Generalized cross correlation with phase transform(GCC-PHAT) [37] is another popular AoA estimation algorithm. It computes the cross-correlation between the received signal and reference signal and estimates the delay to each microphone. Then the delay can be used to search for the best AoA. Increasing the SINR of the received signal improves the accuracy of the cross correlation, which in turn improves the AoA estimation.

Both AoAs and the clean target speech are unknown in the blind speech separation. Thus, we propose to learn the speech and AoA estimation jointly from the acoustic signals.

Learnable Pre-mask [6] propose to incorporate a given AoA into the multi-channel features to guide the network to pay more attention to the TF bins that are dominated by the signal from the given AoA. They develop a pre-mask to take into account AoA $\theta$. The key idea of pre-mask is to compute the correlation between the multi-channel phase representation and the steering vector of AoA $\theta$ across all frequencies and frames. When only target speech exists in the TF bin, the correlation is high. Otherwise, the interference and noise result in a large distortion to the phase.

The pre-mask first forms a steering vector $e_\theta(f)$ based on the AoA, and computes the cosine distance between the steering vector and the complex values in each TF bin as $A(t, f) = \sum_{k=0}^{M} \frac{|w(f, AoA) \cdot Y_k|}{|w(f, AoA)|}$, where $Y$ is the complex spectrogram, $M$ is the number of microphone channels, and $k$ is the microphone index starting from the second microphone as the steering vector is normalized to the first microphone. $A(t, f)$ represents the pre-mask value of a TF bin. The pre-mask indicates the probability of a TF bin dominated by the source coming from the given AoA. Intuitively, the pre-mask lets the network beamform in a given direction. Pre-mask improves over traditional linear beamformers by using a DNN-based non-linear filter, so it has better spatial discrimination and interference cancellation.

The pre-mask assumes that the input AoA of the target user is accurate. In our context, the AoA estimation can be erroneous due to interference, ambient noise, and multipath propagation. Moreover, not only the direct path but also the reflected paths are important for source separation because the overall received phase is the result of all multipath. A single AoA estimate does not provide complete spatial information.
about the target speaker. Therefore, instead of directly using AoA, we propose to fuse our spatial embeddings with the mixed phase from the complex spectrogram to learn a better spatial pre-mask. signals with multichannel phase.

For each microphone and TF bin, there is a 512-long embedding from audible profiles and another 512-long embedding from inaudible profiles. These embeddings are concatenated and processed by 1x1 convolutional layer followed by a layer normalization and ReLU. The output of each TF contains a spatial feature map. It is concatenated with LPS and fed into Temporal Convolution Network (TCN) [13]. TCN outputs a mask, which can be applied to the mixture complex spectrogram to generate the target complex spectrogram. Then we perform an inverse short-term Fourier Transform to estimate the target signals.

**Multi-task Learning Target:** A common learning objective in the existing separation network is to maximize Scale-Invariant Signal-To-Noise Ratio (SiSNR) [13]. Let $\hat{x}$ denote the estimated signal and $x$ denote the clean reference signal. We compute SiSNR as follow:

$$\text{SiSNR} = 10 \log_{10} \frac{\| \hat{x} \|^2}{\| x - \hat{x} \|^2}$$

where $x_{\text{target}} = \frac{x + x_{\text{noise}}}{2}$ and $x_{\text{noise}} = \hat{x} - x_{\text{target}}$. By normalizing $x$ and $\hat{x}$ to zero mean, we ensure scale invariant. The loss function is defined as $L_{\text{SiSNR}} = -\text{SiSNR}$. Following the existing work (e.g., [24]), the target and interference signals are measured separately and added up to simulate interference. Therefore, SiSNR can be computed based on their values.

Unlike the existing works that optimize only SiSNR, we develop a novel multi-task learning framework to jointly learn speech separation and AoA. Our key observation is that spatial embedding can benefit both speech separation and AoA estimation. An accurate embedding enables LSTM to accurately estimate the AoA. It also provides good hints for TF bins, which will be fused with spatial knowledge and target speaker direction.

Another important insight is that jointly learning the separation and AoA can reinforce the network to learn the AoA instead of treating the AoA as the fixed input, which prevents the gradient from propagating back and contributing to the training task. In comparison, when the AoA is set to be learnable together with separation, the mixed phase can contribute to the learning objective. By fusing the phase and embedding, the learned phase is more consistent with the separated source. The phase is represented as a 2D tensor, which represents the phase over different microphones and frequencies. We use more convolutional layers to learn the phase using the following objective for training: $L = L_{\text{SiSNR}} + \lambda L_{\text{AoA}}$ where $\lambda$ is a relative weighting factor and set to 0.5 in our evaluation.

### VII. EVALUATION

#### A. Evaluation Methodology

**Setup:** We connect a laptop with a Bela platform [38] attached with a pair of speakers and four microphones. The microphones form a linear array spanning 8 cm with non-uniform space between them. Their positions are [0, 3 cm, 5 cm, 8 cm]. A similar setup is used in [39], which is feasible to capture the inaudible and audible bands without ambiguity. Certain mobile devices have a similar setup. For example, Apple Macbook Pro [40] places three mics on the right up of the keyboard. Huawei Matebook 14s [41] has a front-facing quad-mic array. Thus, our prototype does not bring extra complexity to the hardware design. The processing pipeline is done on the laptop using Pytorch. We train the model on NVIDIA GTX 2080ti GPU. We run inference on both the desktop and Macbook.

**Data collection:** There are no open-source multi-channel recordings for our evaluation. Therefore, we collect the data on our own. We will release the data to the public. Our data include multi-channel inaudible and audible audio signals. We let dual speakers both transmit periodic FMCW chirps from 18-20 kHz with a period of 40 ms and a sampling rate of 44.1 kHz. To avoid interference, the two speakers transmit the same chirp with 20 ms difference in the starting time. The volume of the speakers is set to a little less than the maximum to avoid signal distortion. We set another headset microphone to record the reference clean speech.

We collect 20 users’ speeches using our setup. Among them, there are 8 females and 12 males. There are two kids: 8 and 13 years old, and the rest are adults between 22 – 59 years old. Each user speaks for 10 minutes - 1 hour. We let the users present slides or read books or papers to mimic online conferences. This is an easy way for users to generate continuous speech. The target user is 0.2–0.7 m away from the microphone. The users move naturally during the trace collection. For example, they sometimes lean towards or away from the computer, move side to side, or turn their heads. We collect the data from various environments (e.g., lab, living room, study room, cubicle, conference room). The environments have different multipath, which affects both AoA estimation and separation.

We also separately record interference by letting an external speaker play a random subset of speech (e.g., around 5 hours) from LibriSpeech [42], which contains more than 1K speakers and 26K English sentences lasting 1000 hours. We place the interfering speaker inside or outside the room where the target user is located. When the interfering speaker is inside the room, (s)he is a couple of meters away. We move the speaker randomly when the speaker plays back the interference. We augment the real traces by scaling the SNR of the target signals from -6 dB to 6 dB. Moreover, we use gpuRIR [43] to simulate realistic interference and noise in multi-channel scenarios by estimating and applying Room Impulse Responses (RIR) to clean speech from Librispeech and noise from WHAM! [44].

**Dataset Preparation:** We mix the audio segments containing the target speaker’s speech and inaudible FMCW reflection with different types of interference and background noise. We add different interference and noise to each target user’s speech. We vary the amount of interference and background noise according to the required SNR. The number of interfering users is uniformly distributed between 0 and 3, and the SNR is uniformly distributed between [-6, 6] dB. In total, the training data is generated from 16 users’ speech. It contains...
30K segments of mixed audio signals, where each segment lasts for 4 seconds and the total training data lasts for 31 hours. The testing data contains 6K segments generated from 4 users. Following the common practice, we vary the user in the testing dataset and use the remaining user for training. Both training and testing have real recording samples from all environments. Interference and noise are from the training split and test split of LibriSpeech and WHAM! respectively. In addition, we also evaluate how our model generalizes to a new environment that is not present in the training traces.

Performance metrics: Following the existing works (e.g., [24]), we use several metrics to quantify the performance of source separation: (i) SiSNR prevents unfair impact of the rescaled signals [45]; (ii) Short-time objective intelligibility measure (STOI) quantifies intelligibility of speech [46]; (iii) Perceptual Evaluation of Speech Quality (PESQ) [47] is designed to quantify the quality of processed speech, and its score ranges from 1 to 5. Higher values in the above metrics indicate better speech quality.

In addition, we report AoA estimation errors. We measure the ground truth AoA using Intel RealSense L515 [48].

Baselines: We compare SAMS with the following state-of-the-art approaches: (i) Conv-TasNet [13]: It is one of the best speech separation approaches using single-channel speech. It is also one of the most widely used baselines due to its open-source. (ii) PHASEN [14]: It is a denoising network using two streams to improve phase estimation. UltraSE [24] shows that [13] and [14] are best baselines that only use speech for source separation. All schemes are trained using the same data.

We did not compare with UltraSE [24], which targets phone users. We target computer users for online meetings, which is complementary to UltraSE. More users use computers than smartphones for their online meetings (e.g., [49] reports the majority use computers for online meetings). Moreover, UltraSE requires the phone’s speaker/mic to face the user’s mouth and be within 20 cm, which is even less common as most speakers/mics on the phone face bottom instead of user. Meanwhile, our measurement shows that the headset Sennheiser DK-2750 improves SiSNR by 8.91 dB over the internal microphones of the laptop under interference. In comparison, our software-only solution provides higher SiSNR, and hence is more attractive.

B. Performance Results

We compare the overall performance of SAMS with several existing source separation methods by varying the background interference, users, SNR, and environments.

Various setups of interference and noise: Following UltraSE [24], we compare our algorithm with Conv-TasNet and PHASEN under different numbers of interfering speakers and noise. All schemes are trained using the same data. Note that SAMS and PHASEN only require the target signal for training, whereas Conv-TasNet requires both the target signal and interference for training. To support multiple interferers, it takes the total interference from all interferers as the ground truth output for training. Table I summarizes the performance in terms of SiSNR, PESQ and STOI. As it shows, our algorithm improves over Conv-TasNet and PHASEN by 5.00 dB, and 1.39 dB, respectively, under only ambient noise. The corresponding numbers become 3.39 dB, and 9.58 dB respectively, under 1 interfering speaker with ambient noise; and become 5.01 dB, and 8.10 dB, respectively, under 2 or more interfering speakers and ambient noise. The larger improvement over the existing approaches under interference is owing to the spatial embeddings learned from both audible and inaudible signals and multi-task learning. Conv-TasNet cannot perform well with only noise or more interference. PHASEN can deal with phase distortion caused by ambient noise, but cannot handle interference well. SAMS can outperform all of them even in their target scenarios. SAMS also achieves better PESQ and STOI as it reduces the phase distortion.

Impact of different multi-channel AoA fusion: There have been various approaches to apply the AoA to help speech separation. We use different AoA fusion strategies in our model, and compare the performance of speech separation. We use the same set of data for training. Note that we still use the TCN separation module and the LSTM AoA module, but we employ the following methods for separation: (i) LPS [50]: It uses only a single channel LPS of raw mixed audio signals for source separation. It uses the same TCN module and does not need any AoA information. (ii) MVDR [19]: We estimate the AoA by applying MUSIC to audible signals, and use MVDR to beamform towards the estimated AoA direction. The beamformed signal is fed to the separation module. (iii) Est AoA pre-mask: We estimate the AoA based on both audible and inaudible features using MUSIC. We use the pre-mask based on the estimated AoA. fuse the pre-mask with LSP, and feed it to the TCN separation module. (iv) GT AoA pre-mask: We directly construct the pre-mask from the ground truth AoA, and apply the pre-mask to separation. (v) Embed: Instead of using the pre-mask, it uses spatial embeddings generated by AoA estimation. However, the AoA estimation module is trained first. Then the weights are frozen during speech separation training. That is, multi-task learning is disabled. (vi) MT: It is the standard SAMS.

As shown in Figure 7, on average, LPS yields 10.26 dB SiSNR by leveraging the magnitude information to learn the acoustic model of speech. MVDR beamforming achieves 10.43 dB SiSNR by leveraging the AoA estimate from audible

| Environment | Model | SiSNR | PESQ | STOI |
|-------------|-------|-------|------|------|
| noise       | SAMS  | 10.71 | 2.21 | 0.76 |
| + interferer| Conv-TasNet | 5.71 | 1.76 | 0.60 |
|            | PHASEN | 9.32 | 2.09 | 0.70 |
| 1 interferer| SAMS  | 13.61 | 2.68 | 0.84 |
| + noise     | Conv-TasNet | 10.22 | 2.19 | 0.76 |
|            | PHASEN | 4.03 | 1.60 | 0.53 |
| 2 or 3 interferers | SAMS | 12.21 | 2.44 | 0.78 |
| + noise     | Conv-TasNet | 7.20 | 1.96 | 0.65 |
|            | PHASEN | 4.11 | 1.69 | 0.53 |

Table I: Compare performance to Conv-TasNet and PHASEN across various interference and noise scenarios.
Table II

quantify the contribution of each component to performance across various interference and noise scenarios

| Environment | Model                      | SiSNR  | PESQ | STOI |
|-------------|----------------------------|--------|------|------|
| noise       | SAMS                       | 10.71  | 2.21 | 0.76 |
|             | SAMS (w/ GT AoA)           | 9.51   | 2.09 | 0.71 |
|             | SAMS (w/o inaudible)       | 10.11  | 2.15 | 0.74 |
|             | SAMS (w/ audible)          | 6.32   | 1.83 | 0.74 |
|             | SAMS                       | 13.61  | 2.08 | 0.84 |
| 1 interferer + noise | SAMS (w/ GT AoA)          | 11.52  | 2.40 | 0.79 |
|             | SAMS (w/o inaudible)       | 12.78  | 2.54 | 0.81 |
|             | SAMS (w/ audible)          | 9.93   | 2.14 | 0.77 |
| 2 or 3 interferers + noise | SAMS (w/ GT AoA)          | 12.21  | 2.44 | 0.78 |
|             | SAMS (w/o inaudible)       | 11.67  | 2.41 | 0.77 |
|             | SAMS (w/ audible)          | 7.01   | 1.92 | 0.62 |

Impact of sensing using inaudible signals: Figure 9 shows that using inaudible signals together with audible signals decreases the AoA error by 1.8° and 1.2° over audible signal based sensing without a mask and with a mask, respectively. The reduced AoA error also translates into improved separation performance. Table II shows that SiSNR decreases 0.6dB, 0.83dB, and 0.54dB without inaudible information for three different noise and interference setups (i.e. under the ambient noise, under 1 interfering speaker with ambient noise, under 2 or more interfering speakers and ambient noise). While inaudible signal-based sensing is useful, it alone (denoted as SAMS (w/o audible)) performs less well. These results confirm combining audible and inaudible signals yields the best performance.

Impact of AoA estimation algorithms: We first compare different variants of our AoA estimation. Figure 9 plots the average AoA estimation error. The basic method is to apply standard wideband MUSIC to estimate the AoA (i.e., applying MUSIC to each frequency band and summing up the results across all bands). We then augment the method with various enhancements. As it shows, each of our enhancements, namely mask, LSTM, inaudible tracking, and multi-task learning, helps improve the AoA error. Using a mask reduces the AoA error by 0.5° – 2.2° across different cases by removing the most noisy TF bins to prevent generating incorrect MUSIC spectrum. Using LSTM brings an additional 1.0° improvement over using the MUSIC profile in a single period since it leverages the inherent temporal locality in the movement. Using inaudible tracking further reduces the AoA error by 1.2° by overcoming audible noise and interference and updating the location more frequently. Finally, multi-task learning improves the AoA by another 0.7° through jointly optimizing the source separation and AoA estimation. Putting everything together, we achieve 3.8° AoA estimation error.

Vary SNR: We vary the SNR of the target user from -6dB to 6dB by scaling the target signal. Each subset of a specific SNR includes all linear combinations of interference and noise. As shown in Figure 8, SAMS outperforms Conv-TasNet and PHASEN in all SNR scenarios by about 3dB and 5dB, respectively. Even for the low SNR case, SAMS can separate the weaker target speech and improve SISNR to 7.09dB, which is sufficient for good audio quality in an online meeting.

VIII. Conclusion

We develop a novel system to combat acoustic interference for online meetings. It advances state-of-the-art in acoustic-based tracking by leveraging both audible and inaudible signals. Moreover, it uses multi-task learning to jointly estimate the AoA and separate the source. Our evaluation shows that our system significantly improves over the state-of-the-art. We believe our work is an important step towards enabling online meetings and classes under interference and noise, which have already been playing a major role in our daily lives. Moving forward, we are interested in exploring other context information to further improve the performance of online meetings.

REFERENCES

[1] S. Pradhan, W. Sun, G. Baig, and L. Qiu, “Combating replay attacks against voice assistants,” Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 3, no. 3, pp. 1–26, 2019.
[2] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, “Audiovisual speech source separation: An overview of key methodologies,” IEEE Signal Processing Magazine, 2014.

[3] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation,” in Proc. of ACM SIGGRAPH, 2018.

[4] K.-S. Lee, “Speech enhancement using ultrasonic doppler sonar,” 2019.

[5] “The overlooked environmental footprint of increasing internet use resources,” Apr. 2021.

[6] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, “Multi-channel overlapped speech recognition with location guided speech extraction network,” in 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 558–565.

[7] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, “Adl-mvdr: All deep l. 2019.

[8] T. Jennung, V. Jayaram, S. Seitz, and I. Kemelmacher-Shlizerman, “The cone of silence: speech separation by localization,” arXiv preprint arXiv:2010.06007, 2020.

[9] R. Nandakumar, S. Gollakota, and N. Watson, “Contactless sleep apnea detection on smartphones,” in Proc. of ACM MobiSys, 2015.

[10] J. Lian, J. Lou, L. Chen, and X. Yuan, “Echospot: Spotting your locations M. Wang, W. Sun, and L. Qiu, “Mavl: Multiresolution analysis of voice via acoustic sensing,” Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 5, no. 3, pp. 1–21, 2021.

[11] M. Wang, W. Sun, and L. Qiu, “Mavl: Multiresolution analysis of voice localization,” in 18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21), 2021, pp. 845–858.

[12] C. Cai, H. Pu, P. Wang, Z. Chen, and J. Luo, “We hear your pace: Passive acoustic localization of multiple walking persons,” Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 5, no. 2, pp. 1–24, 2021.

[13] Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation,” IEEE/ACM transactions on audio, speech, and language processing, 2019.

[14] D. Yin, C. Luo, Z. Xiong, and W. Zeng, “Phasen: A phase-and-harmonics-aware speech enhancement network,” in In Proceedings of AAAI, 2020.

[15] D. Rethage, J. Pons, and X. Serra, “A wavelet for speech denoising,” ICASSP, 2016.

[16] J. R. Hershley, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in Proc. of ICASSP, 2016.

[17] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path mm: efficient long sequence modeling for time-domain single-channel speech separation,” in ICASSP. IEEE, 2020.

[18] C. Sabakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, “Attention is all you need in speech separation,” in ICASSP. IEEE, 2021, pp. 21–25.

[19] J. Benesty, J. Chen, and Y. Huang, “A generalized mvdr spectrum,” IEEE Signal Processing Letters, vol. 12, no. 10, pp. 827–830, 2005.

[20] Y. Xu, Z. Zhang, M. Yu, S.-X. Zhang, and D. Yu, “Generalized spatio-temporal mm beamformer for target speech separation,” arXiv preprint 2101.01280, 2021.

[21] L. Chatterjee, M. Kim, V. Jayaram, S. Gollakota, I. Kemelmacher, S. Patel, and S. M. Seitz, “Clearbuds: Clearbuds: A robust method for speech enhancement,” in Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services, 2022, pp. 384–396.

[22] A. Wang, M. Kim, H. Zhang, and S. Gollakota, “Hybrid neural networks for on-device directional hearing,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 10, 2022, pp. 11421–11430.

[23] T. Barker, T. Virtanen, and O. Delhomme, “Ultrasound-coupled semi-supervised nonnegative matrix factorisation for speech enhancement,” in ICASSP, 2014.

[24] K. Sun and X. Zhang, “Ultrasound: Single-channel speech enhancement using ultrasound,” in Proc. of ACM MobiCom, 2021.

[25] Q. Zhang, D. Wang, R. Zhao, Y. Yu, and J. Shen, “Sensing to hear: Speech enhancement for mobile devices using acoustic signals,” Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 5, no. 3, pp. 1–30, 2021.

[26] C. Xu, Z. Li, H. Zhang, A. S. Rathore, H. Li, C. Song, K. Wang, and W. Xu, “Wavecar: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface,” in Proc. of MobiSys, 2019.

[27] T. Liu, M. Gao, F. Lin, C. Wang, Z. Ba, J. Han, W. Xu, and K. Ren, “Wavecar: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals,” in Proc. of ACM SenSys, 2021.

[28] J. Wang, D. Vaisisht, and D. Katabi, “Rf-idraw: virtual touch screen in the air using rf signals,” ACM SIGCOMM Computer Communication Review, vol. 44, no. 4, pp. 235–246, 2014.

[29] R. O. Schmidt, “A signal subspace approach to multiple emitter location spectral estimation,” Ph. D. Thesis, Stanford University, 1981.

[30] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” IEEE/ACM transactions on audio, speech, and language processing, p. 1849–1858, 2014.

[31] G. Hu and D. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” IEEE Transactions on neural networks, vol. 15, no. 5, pp. 1135–1150, 2004.

[32] Y. Wang, A. Narayan, and D. Wang, “On training targets for supervised speech separation,” IEEE/ACM transactions on audio, speech, and language processing, p. 1849–1858, 2014.

[33] J. Cosentino, M. Pariente, S. Cornell, A. Deler forge, and E. Vincent (2020) Librimix: An open-source dataset for generalizable speech separation.

[34] M. Wax, T.-I. Shan, and T. Kailath, “Spatio-temporal spectral analysis by eigenstructure methods,” IEEE transactions on acoustics, speech, and signal processing, vol. 32, no. 4, pp. 817–827, 1984.

[35] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatio-temporal features with 3d convolutional networks,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.

[36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. of IEEE CVPR, 2016, pp. 770–778.

[37] M. S. Brandstein and H. F. Silverman, “A robust method for speech signal time-delay estimation in reverberant rooms,” in 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1. IEEE, 1997, pp. 375–378.

[38] (2017) Belta platform. https://belta.io.

[39] H. Luo, M. Wang, W. Yu, L. Qiu, S. Pradhan, and Y.-C. Chen, “Rnn-based room scale hand motion tracking,” in The 25th Annual International Conference on Mobile Computing and Networking, 2019, pp. 1–16.

[40] (2021) MacBook tech specs. [Online]. Available: https://www.apple.com/macbook-pro-14-and-16/specs

[41] (2021) Matebook 14s tech specs. [Online]. Available: https://consumer.huawei.com/en/laptops/matebook-14s

[42] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, in (2021) Matebook 14s tech specs. [Online]. Available: https://www.lifesize.com/en/blog/video-conferencing-statistics

[43] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McMahon, D. Crow, E. Manilow, and J. L. Roux, “Wham!: Extending speech separation to harmonics-aware speech enhancement network,” in Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services, 2022, pp. 384–396.

[44] A. Wang, M. Kim, H. Zhang, and S. Gollakota, “Hybrid neural networks for on-device directional hearing,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 10, 2022, pp. 11421–11430.

[45] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, “Audiovisual speech source separation: An overview of key methodologies,” IEEE Signal Processing Magazine, 2014.

[46] T. Barker, T. Virtanen, and O. Delhomme, “Ultrasound-coupled semi-supervised nonnegative matrix factorisation for speech enhancement,” in ICASSP, 2014.

[47] K. Sun and X. Zhang, “Ultrasound: Single-channel speech enhancement using ultrasound,” in Proc. of ACM MobiCom, 2021.

[48] Q. Zhang, D. Wang, R. Zhao, Y. Yu, and J. Shen, “Sensing to hear: Speech enhancement for mobile devices using acoustic signals,” Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 5, no. 3, pp. 1–30, 2021.