VoiceFind: Noise-Resilient Speech Recovery in Commodity Headphones

Irtaza Shahid, Yang Bai, Nakul Garg, Nirupam Roy
{irtaza, yangbai8, nakul22, niruroy}@umd.edu
University of Maryland College Park

ABSTRACT

Robust speech enhancement is a key requirement for many emerging applications. It is challenging to recover clear speech in commodity devices, especially in noisy real-world scenarios. In this paper, we propose *VoiceFind*, which uses only two microphones to spatial filter the desired speech from all interference. Furthermore, to improve the intelligibility of the speech after filtering, we design a Conditional Generative Adversarial Network (cGAN) to reconstruct the desired speech from environmental noises and interference speeches. This is an early attempt to explore this direction. Results from simulation and real-world experiments show promise.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Hardware** → **Beamforming**.

KEYWORDS

Spatial filtering, Direction of arrival, Speech enhancement, Noise-cancellation, Wearables

ACM Reference Format:
Irtaza Shahid, Yang Bai, Nakul Garg, Nirupam Roy. 2022. *VoiceFind*: Noise-Resilient Speech Recovery in Commodity Headphones. In *1st ACM International Workshop on Intelligent Acoustic Systems and Applications (IASA '22), July 1, 2022, Portland, OR, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3539490.3539600

1 INTRODUCTION

Teleconferencing and online audio-video chats have become a part of our daily necessities. Voice communication over the internet connection (VoIP) has made it affordable to the masses and fueled the culture of ubiquitous conversation on mobile devices. The recent pandemic has served as an impetus to the growth of online voice communication. As a result, multiple people conversing on smartphones or headphones are a common sight in homes, in public places, or on daily commutes. This growing culture of voice communication in shared spaces underscores the need for technical innovation in isolating a conversation in a noisy environment. Noise-cancelling earphones [13] solve one-half of the problem by stopping noise while listening. However, while speaking the ambient noise is mixed with the user's voice and degrades its intelligibility to the person at the other end of the conversation. We explore a novel direction of using the frequency structures and unique features of the human voice to recover intelligible speech in extreme noise conditions. This paper presents our speech recovery method that combines ideas in analytical spatial filtering and with generative deep learning approaches. Figure 1 shows an overview of our system.

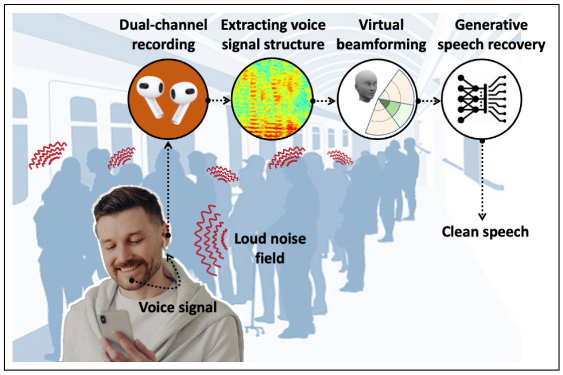

Figure 1: A representative application and overview of *VoiceFind*.

The idea of a spatial filter to suppress noise is not new to the research community. Past works have leveraged spatial filtering to enhance speech coming from a specific angle while suppressing sounds from other directions. Traditional beamforming methods combine multiple spatially distributed sensor data for spatial filtering. Barlett (Delay-Sum) beamforming combines signals with pre-defined delays. Adaptive beamforming such as Minimum Variance Distortionless Response (MVDR), Linear Constraint Minimum Variance (LCMV), and Generalized Side Lobe Canceller (GSC) adaptively tune their filter weights to suppress any signal coming from undesired directions. In the past several years, deep learning-based solutions have demonstrated significant advancement in speech separation. However, these models do not generalize well in high noise scenarios [15]. These solutions also suffer from label permutation problems, which means the model cannot identify the target speech. Instead of only using an audio channel, recently, UltraSE [24] and Hybrid-Beam [26] solve the label permutation problem by capturing the ultrasound reflections from the speaker's face and estimating the angle of arrival with traditional beamforming. These two methods have shown good spatial speech filtering ability, but they rely on multimodal signals or a large microphone array. The requirement of multiple sensors limits their application on commodity devices, such as smartphones and headphones, which commonly have two microphones. To provide a better call experience, the acoustic industry actively explores speech spatial

IASA '22, July 1, 2022, Portland, OR, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9403-1/22/07…$15.00
https://doi.org/10.1145/3539490.3539600

filtering. Sony's voice pickup [2] technology and Apple's voice isolation [1] feature use beamforming microphones and AI-based models to isolate the user's voice. In this paper, we aim to overcome the limitations of training-based speech enhancement and introduce a spatial filtering method based on the frequency structure of the human voice. We believe this approach will complement the existing solutions.

To enable sound separation and enhancement on commodity devices, we ask the question: *Is it possible to spatially filter speech with only two microphones, without any further hardware attachment on commodity devices?* We propose a system, called *VoiceFind*, that can spatially filter human speeches with only two microphones on commodity devices. To achieve this goal, we solve two core challenges: (1) *How to emulate a microphone array with only two of them?* Traditional spatial filtering techniques utilize the phase delay between the recorded signals by microphones caused by propagation in space and time, thus an array is required. To address this challenge, we leverage the fact that human speech includes multiple harmonic frequencies. Instead of using phase difference accumulated in space, we use the phase accumulated in harmonic frequencies to create a modified steering vector. Unlike traditional steering vectors whose resolution depends on the number of sensors and their geometry, our steering vector treats harmonics present in the speech signal as virtual sensors. After creating a steering vector, we use the MUSIC algorithm to compute the direction of arrival (DOA) for all time-frequency components that are corresponding to the speech signal and then based on the estimated DOA we keep the time-frequency component that is coming from the desired direction.

(2) *How to improve intelligibility for the separated speech?* It is known that the amplitude of the time-frequency spectrogram is critical for speech intelligibility. However, spatial filtering only considers the direction of arrival. With intersection points of two speeches in the spectrogram, it is highly possible that some portions in the desired speech also be filtered out. Moreover, environmental non-harmonic noises can also cause an error in spatial filtering. Therefore, we design a conditional GAN (cGAN) to reconstruct the desired speech with the coordination of the spatial filter spectrogram and that of the raw recorded signal. The strategy is the generator in the GAN model learns which portions to keep in the spectrogram and what are the amplitudes in the kept portions, while the discriminator further improves the reconstructed desired speech by analyzing if it is a real or fake pair with the clean desired speech. Moreover, we design a cepstral-based speech filter to remove any non-harmonic and interference noises before applying cGAN.

This project is a work in progress toward a long-term research commitment focused around enhancing voice communication on smart devices. While there are scopes for improvements in signal shaping and optimized deep learning model, this paper shows the possibility of using predictable patterns in human voice for spatial noise elimination and speech enhancement. At this stage of development, *VoiceFind* makes the following contributions:

- We design a spatial filtering technique for human speech that only requires two microphones, which can be applied on any commodity laptops, smartphones, and smartwatches.
- We design a cGAN model to effectively extract only the desired speech from the distorted recorded sound. Furthermore, we apply a cepstral-based speech filter to remove non-speech environmental noises.
- We implement the system and evaluate it in both simulated and real-world environments.

2 CORE INTUITIONS AND PRIMERS

VoiceFind spatially filters human speeches with only two microphones, leveraging the fact that human speech includes multiple harmonic frequencies. In this section, we first introduce traditional DoA estimation using an array, then we highlight the harmonic structure of human speech.

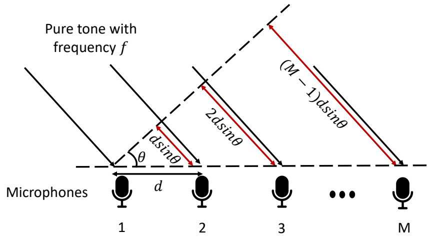

Figure 2: **Direction of arrival estimation using a sensor array.**

2.1 Direction of Arrival Estimation

When a sound wave travels through space and time, it accumulates phase. Computing the phase difference between the waves received at multiple sensors to estimate the Direction of Arrival of a wave is a highly explored technique. Let's say there are M sensors spaced linearly with a distance d apart from each other. There is a sound wave with frequency f traveling at speed c received by these sensors making an angle of θ with the normal to the microphone array. Figure 2 shows the explained setup. It is evident from the Figure 2 that path length is different for each sensor. Path difference L with respect to first element is $L(m) = (m-1)dsin(\theta)$, where m corresponds to the element number in an array ranges from 1 to M. Time delay due to this path difference is $\tau(m) = \frac{L(m)}{c} = \frac{(m-1)dsin(\theta)}{c}$. So, the phase difference ψ between the array elements with respect to the first sensor turns out to be

$$\psi(m) = exp(\frac{j2\pi f(m-1)dsin(\theta)}{c})$$

This equation shows that the phase difference across sensors in an array is a function of angle of arrival θ, frequency of received signal f, and distance between elements in an array d. MUSIC algorithm [20] is one of the most extensively used algorithms in estimating the direction of arrival of the received signal using an array of sensors. The direction is estimated by projecting the signal onto its subspace. Let X be the transmitted signal, and $Y(m)$ be the signal received by the m^{th} sensor. Using the phase difference, we find $Y(m) = exp(j2\pi f(m-1)dsin(\theta)/c)X$, where m ranges from 1 to M corresponds to the element number in an array. This relation treats the phase values as normalized with respect to the first sensor. Defining a Mx1 steering vector $A(\theta)$ in which m^{th} elements is equal to $exp(j2\pi f(m-1)dsin(\theta)/c)$. Now we can write the relation between the transmitted signal and received signal by the sensory array in a compact matrix form as $Y = A(\theta)X$. By finding the eign vectors of Y, and then with the help of a defined steering vector estimate the DOA. The problem setup of the MUSIC algorithm requires having a well-defined steering vector. Moreover, the estimation accuracy is proportional to the number of sensors. Instead of using an array of sensors, we propose a new technique in which multiple frequencies present in the signal are treated as virtual sensors. In this way, we can enjoy the benefit of a large sensory array by using only 2 microphones.

2.2 Structure of Human Speech

Instead of using a microphone array, we seek the opportunity to use the harmonic structure of human speech for DoA estimation. Human speech is produced when air passing through the lungs is modulated by the vocal cords and tracts. The vocal cords vibrate during the pronunciation of a speech which produces voiced signals. Voiced signals have a harmonic pattern in their frequencies like the vowels */a/, /e/, /i/, /o/, /u/*. Unvoiced signals, on the other hand, do not have a harmonic structure as they do not require vocal cords and they are like consonants */f/, /p/, /k/*. Figure 3 shows the spectrogram of recorded sounds of alphabets */f/* and */a/*. We see the harmonic structure of frequencies in the second half of the spectrogram when the user is speaking */a/*. We leverage this harmonic structure to distinguish and filter human speech from noise (explained in detail in section 3.2).

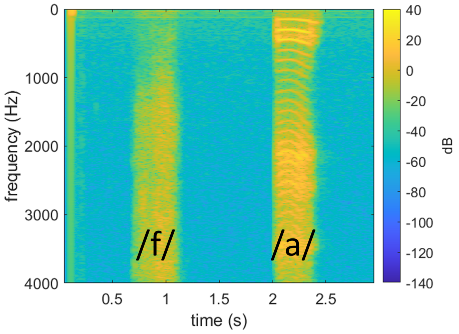

Figure 3: Spectrogram of a speech signal pronouncing /f/ and /a/.

3 SYSTEM DESIGN

The system design of a *VoiceFind* comprises three modules: a) a cepstral-based speech filter to remove any noise and non-harmonic interference by estimating the pitch frequency of the speech signals, (b) a harmonic-based MUSIC algorithm to spatially filter the desired speech using harmonics as a virtual sensors and (c) a conditional GAN to reduce any distortion and generate perceptually pleasant intelligible speech. Next, we will discuss each module in more detail.

3.1 Cepstral-based Speech Filter

The goal of this module is to remove any noise and non-harmonic interference signal. For this purpose, we apply the cepstrum technique which is generally used for pitch estimation of a speech. Cepstrum is the FFT of a log of a signal spectrum, which is used for the analysis of periodic structures in signal spectrums [6]. The frequency of human speech is periodic (has a harmonic structure) with the pitch of the speech signal. Computing FFT of any periodic signal gives the peak at the harmonic frequencies of the periodic signal. We repeat this process for each time window to have pitch frequency estimates with time, we only keep those frequencies and their harmonics. This allows us to keep all speech signals coming from any direction while removing any noise and interference signals. We only pick those peaks lying inside the human speech pitch frequency range [50Hz - 400Hz]. Figure 4 is a spectrogram of a human speech, and on top of that, we have marked pitch frequencies estimated by this module.

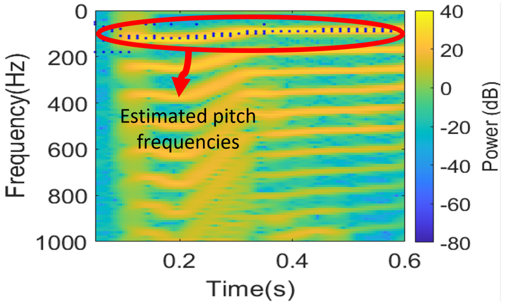

Figure 4: Pitch frequencies estimated by Cepstral-based filter.

3.2 Spatial Filter for Speech

Now that we only have speech signals and their pitch estimates, our next goal is to spatially filter the desired speech coming from 0°. This process comprises two steps: 1) Estimate the DOA for all pitch frequency estimates using the MUSIC algorithm and harmonics as virtual sensors. 2) Apply mask-based filtering using estimated DOAs. Next, we discuss both steps in more detail.

3.2.1 Harmonics as virtual sensors. It is established in Section 2.1 and shown in Figure 5 that the phase difference between sensors is a function of frequency f of a transmitted signal, direction of arrival θ, and spacing d between consecutive elements of sensors in an array. Moreover, the performance of a MUSIC algorithm is proportional to the number of elements in a steering vector which is equal to the number of sensors in an array. So, to achieve reasonable DOA accuracy large array of sensors is required. This paper, on the other hand, proposes a novel steering vector whose length is dependent not on the number of sensors, but the number of harmonics in the signal, by exploiting the fact that phase difference is also a function of the transmitted frequency. Let's say S is a transmitted signal containing N number of different frequencies, and F is an Nx1 vector containing the frequencies present in the signal S. Then $A(\theta)$, the proposed steering vector is defined in such a way that n^{th} element equals to $exp(j2\pi F(n)dsin(\theta)/c)$. Traditional techniques keep the frequency f constant and develop a steering vector by changing distance d, while we keep the distance d constant and alter the center frequency.

The next step is to formulate a measurement matrix so that we can use the MUSIC algorithm for DOA estimation with the proposed steering vector. The development of the measurement matrix is explained below. For a short time window when harmonic frequencies in a speech remain constant, we can represent a signal S in time-domain as $S(t) = \sum_{i=1}^{N} exp(j2\pi F(i)t)$. Its frequency domain representation S(f) is $S(f) = \sum_{i=1}^{N} \delta(f - F(i))$. At the receiver end, we use 2 microphones. After normalizing the phases with reference to the first microphone, the frequency domain of received data of two microphones can be represented as

$$Y1(f) = \sum_{i=1}^{N} \delta(f - F(i))$$

$$Y2(f) = \sum_{i=1}^{N} \delta(f - F(i)) exp(j2\pi F(i) d sin(\theta)/c)$$

Y1 corresponds to first microphone data, and Y2 corresponds to second microphone data. Nx1 measurement vector 'D' is created in such a way that

$$D(i) = \frac{Y2(F(i))}{Y1(F(i))}$$

Now, we can get D as $exp(j2\pi F(i) d sin(\theta)/c)$. This equation is aligned with the problem setup for the MUSIC algorithm. So, now we can perform eigenvalue decomposition and estimate the direction of arrival. In this way, we can estimate DOA using only two microphones with harmonics as virtual sensors. However, this equation only holds when there is no multipath. We argue that the multipath is negligible when the microphones are close to the speaker. The errors caused by multipath in spatial filtering will be recovered by cGAN, which will be introduced in the next section.

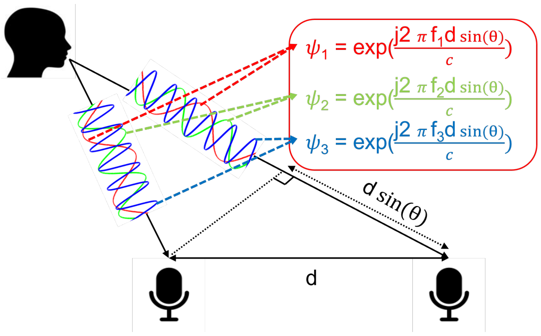
ψ_1 = exp($\frac{j2\pi f_1 d \sin(\theta)}{c}$)
ψ_2 = exp($\frac{j2\pi f_2 d \sin(\theta)}{c}$)
ψ_3 = exp($\frac{j2\pi f_3 d \sin(\theta)}{c}$)
d sin(θ)
d

Figure 5: Frequency-based spatial filter formulates the measurement matrix for the MUSIC algorithm by exploiting the fact that phase difference is also a function of the transmitted frequency.

3.2.2 Binary Mask. To remove all harmonic signals that are not coming from the desired direction, we rely on binary mask B. A binary mask is a binary matrix with a size equal to the speech spectrogram, $B(f,t) = 1 \forall n f_h(t)$, where $f_h(t)$ is the fundamental frequency of harmonics coming from desired direction at time t, and n is any positive integer for which $nf_h(t) < Fs/2$, where Fs is the sampling rate. We estimate the direction of arrival for all fundamental frequency estimates and for all time instances. If an estimated direction is in a pre-defined range of the desired direction, then we consider it as desired harmonic signal and generate our binary mask accordingly. After that, element-wise multiplication of binary mask with received spectrogram returns spatially filtered signal containing only the harmonics from desired speech.

3.3 cGAN-based Speech Enhancement

Now that we have spatially filtered speech signals, we need an enhancement block to improve intelligibility and generate a legible and perceptually pleasant speech. After passing through the cepstral speech filter and spatial filter, the resultant speech still suffers from distortions due to the following reasons: a) the overlapping regions of two speeches get removed by the spatial filter. b) multipath and environmental noise cause some points to be mistakenly included or excluded from the filtered spectrogram. These missing regions and unwanted points induce discontinuity inside the harmonics and spurious frequencies which decreases the intelligibility of the speech. We deploy a cGAN-based speech to correct such distortions and enhancement the speech quality.

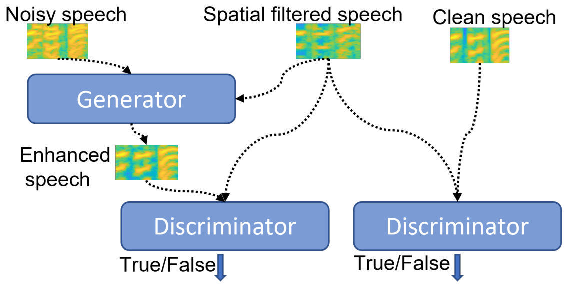

Figure 6: cGAN architecture where discriminator learns classifying between clean and enhanced speech and generator learns beating discriminator by generating enhanced speech similar to the clean speech.

GAN models are explored in detail for speech enhancement and have shown significant advancement over the last couple of years [17]. SEGAN [16] proposed a GAN-based speech enhancement network, and then [17] further improved the performance by proposing two modifications: iterated SEGAN (iSEGAN) and deep SEGAN (DSEGAN). In [23], time-frequency masking-based speech enhancement is proposed using GAN. Daniel et al. [12] have proposed a frequency domain speech enhancement model based on conditional GAN (cGAN) to learn a mapping from the spectrogram of noisy speech to an enhanced counterpart. Therefore, we also adopt a cGAN to further enhance the spectrogram, as shown in Figure 6.

In our model, the generator has two inputs, the spatial filter output, and the raw recorded signal. After the generator learns how to improve the spatial filter output based on the raw recorded signal, i.e., where to add the missing regions and points, the discriminator discriminates whether the enhanced signal matches with the clean desired signal or not. Since the output of the spatial filter is the filtered STFT of the recorded signal, we also use the STFT of the raw recorded signal and desired signal as the input of cGAN. The architecture of our GAN model is similar to the one used by Pix2Pix [8] which has shown great potential in image-to-image translation. Our problem is similar to the image-to-image translation, as we convert a distorted speech spectrogram to a clean speech spectrogram. For the generator, we use an adversarial loss to enforce the generation of realistic spectrograms and L1 loss to enforce the conditional dependence of generated speech spectrogram on input spatially filtered spectrogram. Then we train our network for 200 epochs using raw recorded and spatially filtered spectrograms as input and a clean speech spectrogram as a target. The result in Figure 7 shows that a) the received signal contains two overlapping speech signals, b) spatially filtered spectrogram contains only harmonics corresponding to desired speech but have some distortions, c) cGAN constructed result demonstrates that model is able to learn the harmonic structure, and able to generate realistic speech spectrogram, and d) shows the clean speech for comparison.

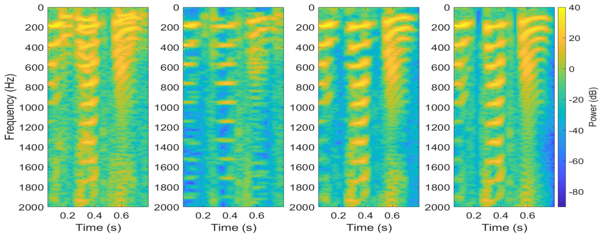

Figure 7: Spectrogram of (a) raw recorded signal (b) spatial filter (c) cGAN reconstructed speech and (d) clean speech.

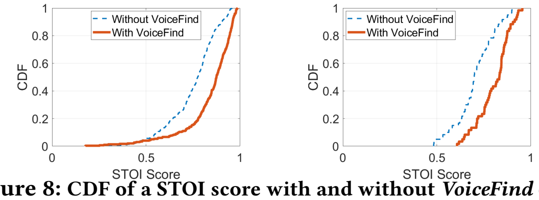

Figure 8: CDF of a STOI score with and without *VoiceFind* on a (a) simulated and (b) real-world collected data.

4 EVALUATION

4.1 Experimental Setup and Dataset

We used two microphones separated by 15cm (the average distance between human ears) for data collection. For data collection, we recorded 400 words in 6 different voices from 0° to 90° with the 10° step. So, we have 2400 recordings from 10 different directions where each recording is of 2 seconds. Then training set is created as follows: For each recording from 0°, we synthetically mix it with a random recording from any direction between 10° to 90°. For the testing set, we repeat the same process but with 50 words that are not present in the training set. The sampling rate of data recording is 16kHz.

4.2 End-to-End Performance

To evaluate the end-to-end performance of *VoiceFind*, we compare the Short-time objective intelligibility measure (STOI) [25], where it ranges from 0 to 1. The higher STOI the better the intelligibility of speech. We compare the STOI with and without *VoiceFind* using simulated data and real-world data. The data without *VoiceFind* includes speeches from all directions, environmental noise, and multipath. As shown in Figure 8, in both simulated and real-world dataset, the median STOI increases around 16%, which means *VoiceFind* does improve the quality of speech by filtering out the interference speech and noises.

4.3 Angle Separation between Speeches

We evaluate the performance of *VoiceFind* with various angle separations between the desired and interference speech. In spatial filtering, we set the angled buffer as 10°, which means when the angle of arrival is within an angle of $-5^\circ \sim 5^\circ$, we treat it as the desired speech. As shown in Figure 10 (a)(b), we find only when the angular separation between two speeches is within 10°, the performance of *VoiceFind* is worse than raw recording. As long as there is a reasonable angle separation between two speeches, our system improves the quality of speech.

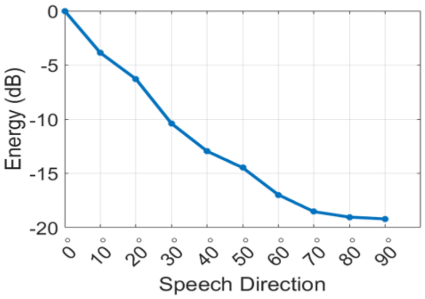

Figure 9: Relative energy in a signal received from different directions after passing through the spatial filter.

4.4 Environmental Noise

We also evaluate the performance of *VoiceFind* under regular kinds of environmental noises, including city traffic, city sidewalk, traffic horns, cafeteria, restaurant, and tone. As shown in Figure 10 (c), the STOI improves around 10% under these environmental noises with *VoiceFind*, meaning *VoiceFind* is effective in filtering out non-harmonic noises.

4.5 Spatial Filter

We evaluate the performance of the spatial filter when the speech comes from directions of 0° to 90°. Ideally only the speech from 0° should be kept, while others are completely filtered out. However, same as traditional DoA estimation techniques, it is not possible to completely filter out the sound from other directions. Each DoA estimation algorithm has a resolution based on parameters such as the number of microphones. To evaluate the resolution of *VoiceFind*, we compute the relative energy after spatial filtering of a signal received from different directions with respect to the signal strength in 0° signal, as shown in Figure 9. We find the signal strength decreases with the increase of the angle, meaning the larger the separation between angles of incoming sound, the better the spatial filter performance.

5 RELATED WORK

Human voice detection and speech recognition using wearable and ubiquitous devices is an active field of research [4, 19, 21]. Estimating the direction of arrival of speech and spatial filtering has also been explored in various contexts [5, 18, 22, 29]. The fundamentals of angle of arrival detection and localization are deeply rooted into the rich literature [9, 11, 28]. MUSIC algorithm [20] is a well explored AoA estimation technique. The basic idea is that when a signal from a propagation path is received across an array of antennas, the AoA introduces a corresponding phase shift across the receivers in the array. While effective, the MUSIC algorithm requires an array for AoA estimation. To improve the accuracy with less antenna, SpotFi [9] uses each frequency component in the WiFi signal to improve the accuracy. There are several existing works that localize human voices [22, 27]. These studies have achieved accurate human voice localization in an indoor environment with strong multipath by re-tracing the paths using the estimated AoA and room structure. Although effective, these works cannot eliminate the interference from another person's voice in the environment.

Deep learning-based solutions have become an integral part of speech separation algorithms. The technique in [7] proposed a

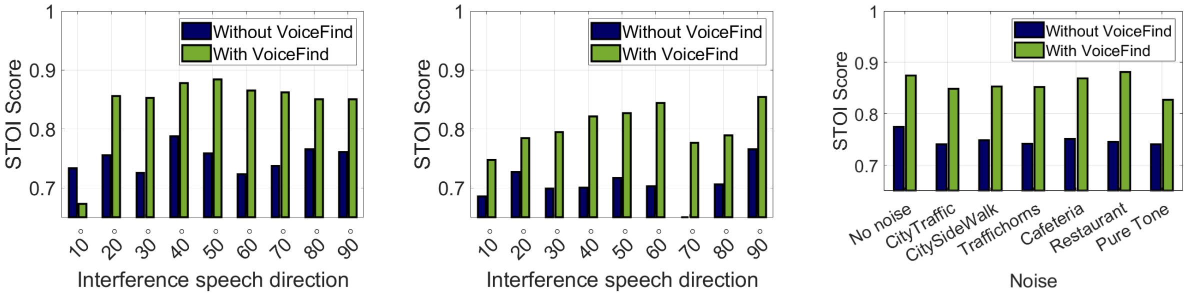

Figure 10: Comparison of STOI score with and without *VoiceFind* on (left) simulated data (center) real data with an interference speech coming from different angles. (right) Median STOI score with and without *VoiceFind* under various noise scenarios.

monaural speech separation technique by jointly optimizing deep neural networks and recurrent neural networks. Conv-TasNet [10] trained a convolutional network to learn the ideal time-frequency magnitude masking for speech separation. Although these solutions have shown remarkable performance in speech separation, these audio-only solutions suffer from label permutation problems. The models cannot identify a target speech. So, the models are trained to separate all superimposed speeches which drastically increases the problem complexity. Recent techniques deal with label permutation problems by gathering information about the target speech using complementary modalities. Audiovisual Zooming [14] leverages video recording of the speaker's face to identify the target speech for separation. But keeping a camera at a certain angle, and under allowable lighting limits its practical usability [3]. Moreover, the recording of a video raises serious privacy concerns. Wang et al. [26] combine traditional beamforming with a neural network to eliminate the interference from a voice in other directions, requiring 6 microphones for the MUSIC algorithm. UltraSE [24] uses ultrasound sensing as a complementary modality to separate the desired speaker's voice from interferences and noise. *VoiceFind* achieves both AoA estimation and spatial filtering of the human voice by using only two microphones.

6 CONCLUSION

In this paper, we use only two microphones to find the direction of arrival of human speech, and spatial filter the desired speech from all interference. We also design a cGAN model to reconstruct human speech after spatial filtering.

REFERENCES

[1] Iphone 13 and iphone 13 mini - technical specifications.

[2] Sony electronics unveils new immersive listening experience with newest industry-leading[1] noise canceling wh-1000xm5 headphones.

[3] Afouras, T., Chung, J. S., and Zisserman, A. My Lips Are Concealed: Audio-Visual Speech Enhancement Through Obstructions. In *Proc. Interspeech 2019* (2019), pp. 4295–4299.

[4] Choudhury, R. R. Earable computing: A new area to think about. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications* (2021), pp. 147–153.

[5] Garg, N., Bai, Y., and Roy, N. Owlet: Enabling spatial information in ubiquitous acoustic devices. In *The 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '21), June 24–July 2, 2021, Virtual, WI, USA*, ACM.

[6] Gudnason, J., and Brookes, M. Voice source cepstrum coefficients for speaker identification. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (2008), IEEE, pp. 4821–4824.

[7] Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. Deep learning for monaural speech separation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2014), IEEE, pp. 1562–1566.

[8] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 1125–1134.

[9] Kotaru, M., Joshi, K., Bharadia, D., and Katti, S. Spotfi: Decimeter level localization using wifi. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication* (2015), pp. 269–282.

[10] Luo, Y., and Mesgarani, N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 27*, 8 (2019), 1256–1266.

[11] McCowan, I. Microphone arrays: A tutorial. *Queensland University, Australia* (2001), 1–38.

[12] Michelsanti, D., and Tan, Z.-H. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. *arXiv preprint arXiv:1709.01703* (2017).

[13] Molesworth, B. R., Burgess, M., and Kwon, D. The use of noise cancelling headphones to improve concurrent task performance in a noisy environment. *Applied Acoustics 74*, 1 (2013), 110–115.

[14] Nair, A. A., Reiter, A., Zheng, C., and Nayar, S. Audiovisual zooming: what you see is what you hear. In *Proceedings of the 27th ACM International Conference on Multimedia* (2019), pp. 1107–1118.

[15] Pandey, A., and Wang, D. On cross-corpus generalization of deep learning based speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 28* (2020), 2489–2499.

[16] Pascual, S., Bonafonte, A., and Serra, J. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452* (2017).

[17] Phan, H., McLoughlin, I. V., Pham, L., Chén, O. Y., Koch, P., De Vos, M., and Mertins, A. Improving gans for speech enhancement. *IEEE Signal Processing Letters 27* (2020), 1700–1704.

[18] Roy, N., and Roy Choudhury, R. Listening through a vibration motor. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services* (2016), pp. 57–69.

[19] Roy, N., Shen, S., Hassanieh, H., and Choudhury, R. R. Inaudible voice commands: The long-range attack and defense. In *15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18)* (2018), pp. 547–560.

[20] Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation 34*, 3 (1986), 276–280.

[21] Sehgal, A., and Kehtarnavaz, N. A convolutional neural network smartphone app for real-time voice activity detection. *IEEE Access 6* (2018), 9017–9026.

[22] Shen, S., Chen, D., Wei, Y.-L., Yang, Z., and Choudhury, R. R. Voice localization using nearby wall reflections. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (2020), pp. 1–14.

[23] Soni, M. H., Shah, N., and Patil, H. A. Time-frequency masking-based speech enhancement using generative adversarial network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), pp. 5039–5043.

[24] Sun, K., and Zhang, X. Ultrase: single-channel speech enhancement using ultrasound. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking* (2021), pp. 160–173.

[25] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing* (2010), IEEE, pp. 4214–4217.

[26] Wang, A., Kim, M., Zhang, H., and Gollakota, S. Hybrid neural networks for on-device directional hearing. *arXiv preprint arXiv:2112.05893* (2021).

[27] Wang, M., Sun, W., and Qiu, L. {MAVL}: Multiresolution analysis of voice localization. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)* (2021), pp. 845–858.

[28] Xiong, J., and Jamieson, K. {ArrayTrack}: A {Fine-Grained} indoor location system. In *10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)* (2013), pp. 71–84.

[29] Zheng, S., Zhang, S., Huang, W., Chen, Q., Suo, H., Lei, M., Feng, J., and Yan, Z. Beamtransformer: Microphone array-based overlapping speech detection. *arXiv preprint arXiv:2109.04049* (2021).