IEEE INFOCOM 2022 - IEEE Conference on Computer Communications | 978-1-6654-5822-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/INFOCOM48880.2022.9796768

# Separating Voices from Multiple Sound Sources using 2D Microphone Array

Xinran Lu[†], Lei Xie[†], Fang Wang[†], Tao Gu[‡], Chuyu Wang[†], Wei Wang[†], Sanglu Lu[†]
[†]State Key Laboratory for Novel Software Technology, Nanjing University, China
[‡]School of Computing, Macquarie University, Australia
luxinran@smail.nju.edu.cn, lxie@nju.edu.cn, mf20330077@smail.nju.edu.cn, tao.gu@mq.edu.au,
{chuyu,ww,sanglu}@nju.edu.cn

***Abstract*—Voice assistant has been widely used for human-computer interaction and automatic meeting minutes. However, for multiple sound sources, the performance of speech recognition in voice assistant decreases dramatically. Therefore, it is crucial to separate multiple voices efficiently for an effective voice assistant application in multi-user scenarios. In this paper, we present a novel voice separation system using a 2D microphone array in multiple sound source scenarios. Specifically, we propose a spatial filtering-based method to iteratively estimate the Angle of Arrival (AoA) of each sound source and separate the voice signals with adaptive beamforming. We use *BeamForming-based cross-Correlation (BF-Correlation)* to accurately assess the performance of beamforming and automatically optimize the voice separation in the iterative framework. Different from cross-correlation, *BF-Correlation* further performs cross-correlation among the *after-beamforming voice signals* processed with each linear microphone array. In this way, the mutual interference from voice signals out of the specified direction can be effectively suppressed or mitigated via the spatial filtering technique. We implement a prototype system and evaluate its performance in real environments. Experimental results show that the average AoA error is 1.4 degree and the average ratio of automatic speech recognition accuracy is 90.2% in the presence of three sound sources.**

## I. Introduction

**Motivation:** Voice assistant has been widely used in a diverse range of application scenarios from human-computer interaction to automatic meeting minutes. Popular voice assistants include Amazon Echo [1], Apple HomePod [2] and Google Home [3], and they utilize speech recognition techniques to improve human-computer interaction. Voice assistants have also been deployed in meeting rooms to record the voice of participants, as shown in Fig. 1. By applying natural language processing (NLP) techniques, meeting minutes can be automatically recorded by translating audio into text. Moreover, localization [4–12] can be further applied to the voice assistant to provide location context to voice recognition. These techniques perform well in the situation of a single sound source. However, in the situation of multiple sound sources, the performance of speech recognition and localization decreases dramatically. For example, as shown in Fig. 1, multiple persons in the meeting room are speaking simultaneously, mixed voice signals are collected by microphones on the table. Due to mutual interference in mixed voice signals, the system is unable to effectively perform speech recognition. Separating voices from multiple sound sources is crucial to ensure the performance of speech recognition.

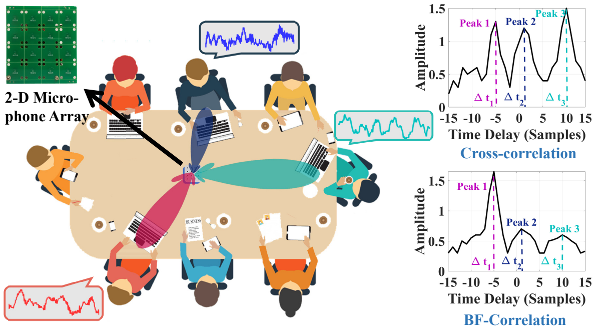


Fig. 1: Voice separation of multiple sound sources via microphone array.

**Limitation of Prior Art:** There are three main approaches to perform voice separation from the mixed voice signals of multiple sound sources. *The first approach is to utilize massive directional microphones to record voice signals*. The directional microphones are physically deployed to record voice signals with a range limit. Each directional microphone only records the voice of the sound source in interest. Voice signals from multiple sources are naturally separated from each other. However, this approach usually requires specific deployment of directional microphones for multiple users. This greatly increases the deployment cost and hardware cost for voice separation. *The second approach is to use signal processing algorithms to separate voice signals*. Independent component analysis (ICA) is a typical computational method for signal separation. However, this approach usually works well in ideal situations with strong assumptions, e.g., the source signals are independent of each other. However, the multi-path effect may introduce large amounts of virtual sources and the signals are correlated with each other. Thus, this dramatically degrades the performance of voice separation in real environments. *The third approach is to use spatial filters such as beamforming to separate voice signals*. This approach separates voice signals according to the spatial position of sound sources, i.e., angle of arrival (AoA) , by adjusting directional antenna gain. Thus, the multi-path effect can be partly suppressed. However, the

978-1-6654-5822-1/22/$31.00 ©2022 IEEE

position or angle of sound sources are usually required to be known in advance. Due to the superposition and mutual interference of voice signals, the traditional AoA estimation methods cannot precisely estimate the angle of each sound source. This greatly limits the applicability of the spatial filter-based approach.

**Our Approach:** Multiple voice signals usually come from different spatial direction since the sound sources, i.e., human voice, are usually separated physically. In this paper, we propose a novel spatial filtering-based technique with iterative AoA estimation and beamforming using a 2D microphone array-based system for voice separation in multiple sound source scenarios. We first perform *cross-correlations* between raw voice signals to get the coarse-grained *angles of arrival (AoA)* of sound sources. Then, according to each estimated AoA for the sound source candidates, we use *beamforming* at receiving ends, i.e., microphone array, respectively. After that, we propose *an iterative approach* to optimize the AoA estimations and the beamforming result. Specifically, during the iteration, we propose the *BeamForming-based cross-Correlation (BF-Correlation)* to further refine AoA estimation. Different from cross-correlation, *BF-Correlation* further performs cross-correlation among the *after-beamforming voice signals* processed with each linear microphone array. By applying the spatial filtering technique, the mutual interference from the signals out of the specified AoA can be effectively suppressed or mitigated in computing *BF-Correlation*. In this way, more refined AoA estimation can be obtained by referring to *BF-Correlation*, which further supervises the subsequent beamforming.

**Challenges:** There are two challenges to be addressed in this paper. The first challenge is to accurately estimate the angle of arrival (AoA) of multiple sound sources from mixed voice signals. Due to superposition and mutual interference of multiple signals, cross-correlation among mixed raw signals usually leads to *Peak Confusion*. Besides, human voices are broadband signals which are partly self-correlated, this further leads to *Fake Peak* in cross-correlation. Therefore, traditional AoA estimation approaches based on cross-correlation, such as GCC-PHAT[13], may not work effectively in the situation of multiple sound sources. To address this challenge, we propose a novel *BF-Correlation*-based scheme for accurate AoA estimation. This scheme is able to obtain a fine-grained AoA estimation by iteratively performing AoA estimation and beamforming at the receiving end, i.e., microphone array. For the 2D microphone array, by performing beamforming towards a specific angle over each row/column of linear array, we are able to suppress or mitigate most of the interference of the voice signals from the other angles. Then, by performing *BF-Correlation* among these *after-beamforming signals*, the peaks in *BF-Correlation* can depict the corresponding time delay of AoA in a more refined manner. The following beamforming can further tune the receive gain towards the refined angle to iteratively achieve better performance. In this way, we can sufficiently reduce the possibilities of peak confusion in cross-correlation.

The second challenge is to automatically optimize the effect of voice separation in the iterative framework. Note that we utilize beamforming to obtain the separated voice signals via the spatial filtering techniques. However, since we do not have prior knowledge of the sound sources, i.e., the ground-truth features of each original voice signal cannot be obtained, during the iteration process, we cannot effectively evaluate the quality of separated voice signals by comparing with the ground-truth, in terms of signal to noise ratio (SNR) and MOS[14]. Hence, a general indicator is necessarily required to perform automatic optimization in the iterative approach. To address this challenge, we use *relative peak amplitude* in *BF-Correlation* to evaluate the effect of voice separation. We find that the peak in cross-correlation usually denotes a candidate sound source with a specific time delay in the microphone array, corresponding to a certain angle. The amplitude of the peak further depicts the strength of voice signals from a certain angle. Since we use beamforming to increase the gain of voice signals from a specified angle, and suppress voice signals from the other angles. Therefore, when the beamforming targets towards the right direction of voice signal, the amplitude of the peak corresponding to the beamforming angle should be much larger than the other amplitudes of the wave. Hence, leveraging the *relative peak amplitude* in *BF-Correlation*, i.e., the ratio of the peak's amplitude to the average amplitude of the wave corresponding to all directions, we are able to evaluate the effect of voice separation with spatial filtering.

**Contributions:** This paper makes the following contributions. First, we present a novel 2D microphone array-based system to perform voice separation in multiple sound source scenarios, by precisely estimating the AoA of each sound source and separate the signals with adaptive beamforming. Second, we propose a novel BeamForming-based cross-Correlation (BF-Correlation) scheme for accurate AoA estimation and adaptive beamforming. We analyze the *peak confusion* issues for *cross-correlation* via thorough empirical study and modeling, and propose *BF-Correlation* to accurately estimate the AoA of multiple sound sources and automatically optimize voice separation in our iterative framework. Third, we implemented a prototype system and evaluate the performance in real environments. Experiment results show that the average AoA error is 1.4° and the average ratio of automatic speech recognition accuracy is 90.2% in the presence of three sound sources.

## II. Related Work

**Voice Signals Separation.** Besides separating the signals from the sources physically, the approaches to separate the mixed voice signals can be divided into two categories. The first category is to utilize spatial filters to separate the voice signals. Early works in this category need to have the priori knowledge of the direction of sound sources to separate the mixed voice signals [15–18]. Some approaches are proposed to estimate the angle of arrival (AoA) for spatial filters, such as GCC-PATH [13], *Multiple Signal Classification* (MUSIC) algorithms [19–21], etc. However, due to self-correlation and broadband characteristics of voice signals, these approaches cannot precisely locate the sound sources in real scenarios. Besides, visual signals are utilized in recent works [22–25]. After locating the sources with extra visual devices, i.e., cameras, spatial filter based approaches are applied to separate the voice signals. The second category is to utilize the characteristics of signals for separation. Traditional approaches in this category are signal processing methods, such as *Independent Component Analysis* (ICA) [26–28]. However, the signal processing methods have strong assumptions. The signals should be independent to each other and the solution should be performed in the scenario without multi-path effect. Thus, traditional signal processing approaches are not suitable for most practical use. Instead of separating all the signals, some recent work extract signals from target speakers [29, 30], with the priori knowledge of the signals, e.g., the frequency patterns or the contents of voice signals.

**Techniques based on Microphone Arrays.** The microphone arrays are recently utilized for signals processing and sensing. Noise reduction [31, 32], speech enhancement [33–35] and other signal processing approaches are conducted with microphone arrays. Moreover, microphone arrays can be used as fine-grained sensors. VoLoc [4] and Symphony [5] use wall reflection and angle of arrival (AoA) algorithms to locate the sound sources via microphone arrays. Poozesh [36] uses a microphone array to monitor the structural health with non-contacting measurement technique. AcuTe [37, 38] uses microphones to sense the temperature of the environment. VSkin [39] uses microphones as motion sensors to control the mobile phones. All these works use the spatial structure of the microphone array. However, voice separation in real scenarios have not been effectively solved in recent works.

## III. Empirical Study and Modeling

### A. Problem Formulation

In this paper, we address the problem of accurate AoA estimation and signal separation of multiple sound sources via a 2-D uniform rectangle microphone array. The basic idea for voice separation is to first utilize cross-correlation between signals from microphones to estimate the AoAs of sound sources. We then apply beamforming methods to separate the voice signals according to the AoAs estimated. Suppose there are $n$ sound sources denoted as $S_1$ to $S_n$, and $k$ microphones denoted as $M_1$ to $M_k$. The signals emitted from $S_1, S_2, \cdots, S_n$ are denoted as $s'_1(t), s'_2(t), \cdots, s'_n(t)$ and the signals received by $M_1, M_2, \cdots, M_k$ are denoted as $m_1(t), m_2(t), \cdots, m_k(t)$. Signals from different sound sources arrive at each microphone at different time. Suppose $\rho_{i,j}$ is the propagation coefficient and $\tau_{i,j}$ is the propagation time from $S_i$ to $M_j$. Then the signal $m_j$ received by $M_j$ is:

$$m_j(t) = \sum_{i=1}^{n} \rho_{i,j} s'_i(t + \tau_{i,j}), \tag{1}$$

Due to the far field propagation model [5], the propagation coefficients from the same sound source $S_i$ are equal for all $k$ microphones. That means the propagation coefficient $\rho_{i,j}$ can be simplified as $\rho_i$. If we use $M_1$ as reference, let $\Delta t_{i,j} = \tau_{i,j} - \tau_{i,1}$ denote the time difference of arrival between signals arriving at $M_j$ and $M_1$ from $S_i$. Let $s_i(t) = \rho_i s'_i(t + \tau_{i,1})$ denote the signals arriving at $M_1$ from $S_i$. The equation is simplified as:

$$m_j(t) = \sum_{i=1}^{n} \rho_i s'_i(t + \tau_{i,1} + \tau_{i,j} - \tau_{i,1}) = \sum_{i=1}^{n} s_i(t + \Delta t_{i,j}). \tag{2}$$

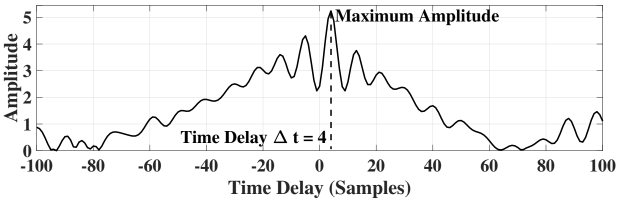


Fig. 2: Cross-correlation for single sound source.

The following part shows the formulation of beamforming. To clearly explain the formulation, we take the classic beamforming, *Delay-and-Sum Beamforming*, as an example. Suppose $\Delta t'_j$ is the time shift of signals received by $M_j$, the outcome of beamforming $m'(t)$ is:

$$m'(t) = \frac{1}{k}\sum_{j=1}^{k} m_j(t - \Delta t'_j) = \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{n} s_i(t + \Delta t_{i,j} - \Delta t'_j). \tag{3}$$

In Eq. (3), we can see if we want to perform beamforming on $S_i$, we need to find the time shift $\Delta t'_j$ to compensate the time delay $\Delta t_{i,j}$. As a result, the goal of the algorithm is to precisely estimate the time delay $\Delta t_{i,j}$ and perform beamforming for each sound source $S_i$.

### B. Modeling Cross-correlation of Multiple Sound Sources

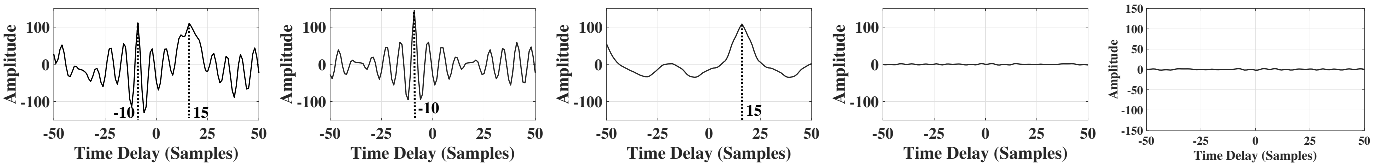
Cross-correlation is a function of displacement of one relative to the other, measuring the similarity of two series[40]. Cross-correlation between $s_i(t)$ and $s_j(t)$ is defined as:

$$Cor_{s_i,s_j}(n) = \sum_{m=1}^{N} s_i(m-n)s_j(m), -(N-1) \le n \le N-1, \tag{4}$$

which is equivalent to convolution of $s_i(-t)$ and $s_j(t)$. $s_i(t)$ and $s_j(t)$ are discrete functions in the field of real numbers with $N$ samples.

*1) Cross-correlation between signals from Single Sound Source:* Suppose we have a microphone array with two microphones denoted by $M_i$ and $M_j$, respectively. The signals received by the two microphones is denoted by $m_i(t)$ and $m_j(t)$, respectively. The cross-correlation between $m_i(t)$ and $m_j(t)$ can be calculated according to Eq. (4). Suppose the single sound source $S$ with signal $s(t)$ is in a free-space without multipath effect and attenuation, the two microphones receive the signals with time delay $\Delta t$, which gives $m_i(t) = s(t)$ and $m_j(t) = s(t + \Delta t)$. There is only one main peak in the cross-correlation figure between signals from the microphones. The time shift $\Delta t$ can be calculated from the correlation with $\Delta t = \arg\max_t Cor_{m_i,m_j}(t)$. Fig. 2 shows the cross-correlation between the received signals in the single-source scenario. The peak of the maximum amplitude corresponds to the time delay $\Delta t = 4$ samples. Suppose the speed of sound in the air is $v$, the distance between the two microphones is $d$ and the sampling rate is $f_s$, the angle of arrival (AoA) $\theta$ of $s(t)$ is:

$$\theta = \arcsin\frac{\Delta t \cdot v}{f_s \cdot d}. \tag{5}$$



(a) Cross-Correlation between $m_1(t)$ and $m_2(t)$.

(b) Cross-Correlation between $s_A(t)$ and $s_A(t+\Delta t_A)$.

(c) Cross-Correlation between $s_B(t)$ and $s_B(t+\Delta t_B)$.

(d) Cross-Correlation between $s_A(t)$ and $s_B(t+\Delta t_B)$.

(e) Cross-Correlation between $s_B(t)$ and $s_A(t+\Delta t_A)$.

Fig. 3: Illustration of Correlation Distribution Property.

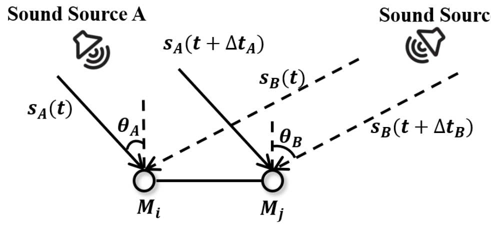


Fig. 4: Multiple sound sources model.

TABLE I: Correspondence between $\Delta t$ and Incident Angle.

| Time Delay (Samples) | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Incident Angle (°) | 0 | 5.7 | 11.4 | 17.3 | 23.4 | 29.8 | 36.5 | 44 | 52.6 | 63.3 | 83.0 |

*2) Cross-correlation between Mixed Signals from Multiple Sound Sources:* For multiple sound sources, the cross-correlations between received signals are different from that in single-source scenario. Suppose we calculate the cross-correlation between $m_j(t)$ and $m_l(t)$, the signals received by $M_j$ and $M_l$ respectively. According to Eq. (2), $m_j(t) = \sum_{i=1}^{n} s_i(t+\Delta t_{i,j})$ and $m_l(t) = \sum_{i=1}^{n} s_i(t+\Delta t_{i,l})$. Since convolution has the distributive property, it is proved that cross-correlation have the distributive property. Thus, the cross-correlation between $m_j(t)$ and $m_l(t)$ is:

$$Cor_{m_j(t),m_l(t)} = \sum_{i=1}^{n}\sum_{k=1}^{n} Cor_{s_i(t+\Delta t_{i,j}),s_k(t+\Delta t_{k,l})}. \qquad (6)$$

We take two sound sources $S_A$ and $S_B$ as an example. As shown in Fig. 4, the received signals $m_i(t)$ and $m_j(t)$ consist of two components $s_A(t)$ and $s_B(t)$ from $S_A$ and $S_B$:

$$\begin{aligned} m_i(t) &= s_A(t) + s_B(t), \\ m_j(t) &= s_A(t+\Delta t_A) + s_B(t+\Delta t_B). \end{aligned} \qquad (7)$$

The cross-correlation between $m_i(t)$ and $m_j(t)$ is:

$$\begin{aligned} Cor_{m_i,m_j} =& Cor_{s_A(t),s_A(t+\Delta t_A)} + Cor_{s_A(t),s_B(t+\Delta t_B)} \\ &+ Cor_{s_B(t),s_A(t+\Delta t_A)} + Cor_{s_B(t),s_B(t+\Delta t_B)}. \end{aligned} \qquad (8)$$

From Eq. (8), we can see that the cross-correlation consists of 4 components, and the components can be divided into two categories: correlation between signals from the same sound source and correlation between signals from different sound sources. Obviously, $Cor_{s_A(t),s_A(t+\Delta t_A)}$ and $Cor_{s_B(t),s_B(t+\Delta t_B)}$ are the correlations between the same sound sources and there exist peaks that indicate two time delays. $Cor_{s_A(t),s_B(t+\Delta t_B)}$ and $Cor_{s_B(t),s_A(t+\Delta t_A)}$ are the correlations between different sources, and since correlation describes the similarity of different signals, the correlation amplitudes are significantly smaller than former correlation peak amplitudes.

We conduct experiments to illustrate the cross-correlation distribution property in multiple sound sources scenario. The experimental settings are described as follows. We deploy two microphones $M_1$, $M_2$ and two sound sources $S_A$, $S_B$ following the layout shown in Fig. 4. The distance $d$ between the two microphones is 15 cm, and incident angle $\theta_A = -30°$, $\theta_B = 50°$, respectively. The distances $l_A, l_B$ between the sound sources and the array are 1 m and 1.5 m. According to Eq. (5), the time delay $\Delta t_A$ should be $-10$ samples and $\Delta t_B$ should be 15 samples. The signals emitted by sound sources are human voice signals. We have two findings from Fig. 3. First, Fig. 3(b) – 3(c) show that the amplitude of correlation between the same source are much larger than that between different sources. Second, the peaks in Fig. 3(a) essentially indicate the time delay $\Delta t_A, \Delta t_B$, respectively, in received signals $m_1(t)$ and $m_2(t)$.

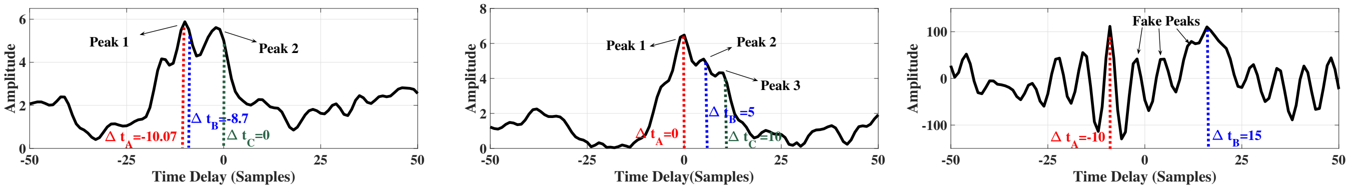
In summary, for multiple sound sources, cross-correlation between signals from different microphones is composed of cross-correlations between different components of received signals. The amplitude of correlation from the same sound source is much larger than that from different sound sources. That means in most cases, peaks in the cross-correlation of received signals imply the corresponding time delay of signals and this makes AoA estimation in multiple sound sources scenario possible. However, we find that there still exist some challenging issues for us to use cross-correlation. The following sections show the issues in cross-correlation and analyze the reasons for them.

### C. Peak Confusion in Cross-correlation

Ideally, AoA can be estimated through peaks in cross-correlation between received signals. However, due to insufficient sampling rate and self-correlated characteristic of human voice, the peaks do not always denote AoAs of the signals. We conduct an empirical study to show possible situations of *Peak Confusion* and further analyze the reasons for *Peak Confusion*.

As shown in Fig. 6, we use $4 \times 4$ uniform rectangle array (URA) with 16 microphones and the distance between the adjacent microphones $d$ is 2.4 cm. The microphones are denoted as $M_1$, $M_2, \cdots, M_{16}$. We build a polar coordinate system for the microphone array. Pole $O$ of the coordinate is located at array center and the polar axis points to the right. There are three sound sources $S_A, S_B, S_C$ emitting three different voice signals. The coordinate of $S_A$, $S_B$ and $S_C$ are $(2m, 180°)$, $(2m, 210°)$ and $(1.5m, 270°)$. The sound sources are in the far field of microphone array and the sampling rate $f_s$ is set to 48 kHz.

We identify three issues, including *peak overlap*, *peak deviation* [5] and *fake peak*, which are observed when performing cross-correlation. *Peak overlap* represents peaks of the ground truths are too close each other, resulting in being difficult to



(a) *Peak Overlap*. Peak 1 is caused by $S_A$ and $S_B$. The time delay of Peak 1 is -10 samples. Peak 2 is caused by $S_C$. The time delay of Peak 2 is -2 samples.

(b) *Peak Deviation*. Peak 1,2,3 are caused by $S_A$, $S_B$, $S_C$, respectively. The time delay of Peak 1,2,3 are 0,4,9 samples, respectively.

(c) *Fake Peak*. The time delays of groundtruth is -10 samples and 15 samples. Several peaks do not refer to any sound source.

Fig. 5: *Peak Confusion* in cross-correlation.

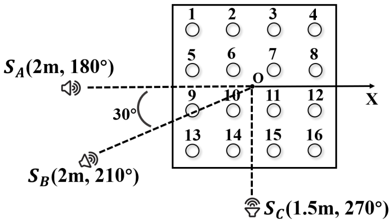
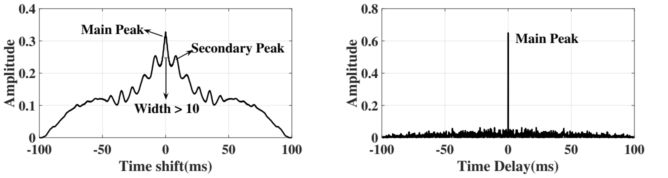


Fig. 6: Settings of empirical study.



(a) Self-correlation of voice signals.

(b) Self-correlation of random signals.

Fig. 7: Self-correlation of different kinds of signals.

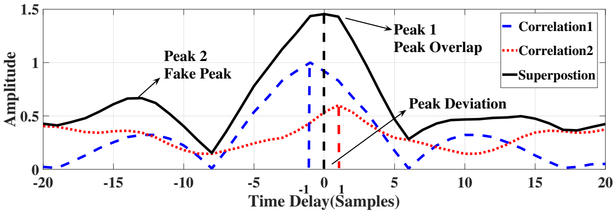


Fig. 8: Superposition of different correlations.

separate. Fig. 5(a) shows the cross-correlation between $m_1(t)$ and $m_4(t)$ from $M_1$ and $M_4$. According to the ground truth, time delays of $S_A$ and $S_B$ should be $-10$ samples and $-9$ samples, respectively. However, there is only one peak in the corresponding region. That means the peak caused by $S_A$ overlaps the peak caused by $S_B$. *Peak deviation* represents that some peaks may be several points away from the ground truth. Fig. 5(b) shows the cross-correlation between $m_1(t)$ and $m_{13}(t)$ from $M_1$ and $M_{13}$. The time delay of Peak 3 is 9 samples and the incident angle we calculate from time delay is about $64^\circ$ ($244^\circ$ in polar coordinate system). The estimated angle deviates dramatically from the ground truth $270^\circ$. *Fake peak* represents that peaks in the cross-correlation do not refer to any sound source. Fig. 5(c) shows the cross-correlation between the signals in Fig. 3. There are three fake peaks in the correlation figure since there are only two sound sources in the scenario and the corresponding time delay is -10 samples and 15 samples, respectively.

There are three reasons for the issues. *First, the spatial resolution of the microphone arrays are limited* [5]. There are 21 candidate bins in cross-correlation for the empirical study as shown in Table I. The peaks overlap each other if the AoAs are located in the same or adjacent bins. *Second, the self-correlation characteristic of the human voices may cause the issues.* Human voices are partly self-correlated, which is different from signals in channel impulse response (CIR) or random signals. In Fig. 7(b), random signals or impulse signals have one main peak. The peak is narrow and easy to recognize. However, the self-correlation figures of voice signals have one main peak and several secondary peaks, and the width of the peaks usually larger than 2 sample points, as shown in Fig. 7(a). Secondary peaks can easily be ignored in the scenario of single sound source while it is hard to recognize in the scenario of multiple sound sources. *Fake peaks* are mainly caused by self-correlation property of voice signals. *Third, the superposition of different correlation affects the peak amplitude and position.* From Eq. (8) we can see that the cross-correlation of signals from multiple sound sources is superimposed by several parts. The peaks caused by different sound sources interfere with each other. Fig. 8 illustrates how superposition affects the peaks in cross-correlation. In the figure, Peak 2 is a *Fake peak* mainly caused by self-correlation property of signal 1. The superposition of main peak of signal 1 (time delay = -1 sample) and main peak of signal 2 (time delay = 1 sample) forms a new peak (time delay = 0 sample). This leads to *Peak overlap* and *Peak deviation*.

Due to *Peak overlap, Peak deviation and Fake peak*, the peaks in cross-correlation cannot always denote AoA of the signals from multiple sound sources. We observe that the effect of beamforming is related to the AoA of different signals. In the meantime, cross-correlation can depict the relative amplitude between different signals. We can handle the issues with iterative framework of correlation and beamforming.

## IV. System Design

### A. System Overview

Fig. 9 gives an overview of the system. The system first receives and segments the raw voice signals. After preprocessing, the system performs *Cross-correlation* in different dimensions, for example columns and rows, and decides which dimension of the microphone array is used for beamforming. Moreover, *Cross-correlation* generates coarse-grained AoA information for beamforming, which points to the potential sound sources. In *Beamforming*, the system first adjusts the array to the direction of potential sound sources, and then performs adaptive beamforming. In *BF-Correlation Assessment*, BeamForming based cross-Correlation (*BF-Correlation*) is performed between outputs of beamforming. The ratios of amplitudes in the direction of potential sound sources

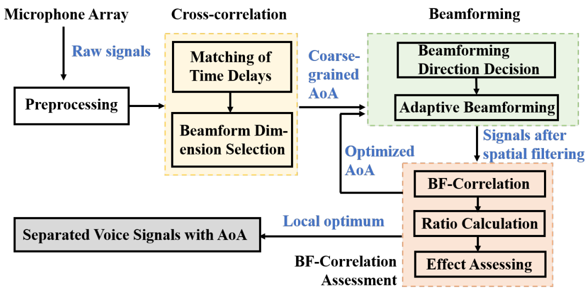


Fig. 9: System Overview.

to amplitudes in other directions is calculated from *BF-Correlation*. The ratios are indicators of beamforming effect. The iteration of *Beamforming* and *BF-Correlation Assessment* provides relationships between optimized AoAs and indicators. Local Search is performed to find the optimum AoAs with the indicators. Finally, the system outputs the separated voices with AoAs.

### *B. Preprocessing*

The voice signals received by the microphone arrays are split into small segments for the subsequent cross-correlation and beamforming, so as to reduce computation and exclude instant noises. The size of segmentation needs consideration. The increase of segmentation size increases the computation and reduces the real-time capability of the system. However , the larger the size of segmentation is, the better the performance of cross-correlation and beamforming will be. To achieve a balance between performance and time consumption, we let one segment last for 2 s.

### *C. Cross-correlation Analysis with 2-D Microphone Array*

*1) Advantages of 2-D Microphone Array:* Traditional approaches using uniform linear array (ULA) cannot precisely estimate AoA of signals for multiple sound sources. Moreover, ULA can only distinguish different AoAs in 180°. In our system, we utilize uniform rectangle array (URA) as shown in Fig. 6. The advantages of utilizing a 2-D microphone array are two-fold: First, 2-D microphone array may perform AoA and beamforming algorithms on sound sources from 360°. Second, 2-D microphone array provides multiple dimensions in the sound source plane, i.e., the row array and column array, to estimate AoA of the signals. We can select a proper dimension from the 2-D microphone array for beamforming and another for evaluation.

*2) Matching Time Delays in Different Dimensions:* To decide which dimension is suitable for beamforming or evaluation, we first analyze the relationship between the time delay between different dimensions. Without loss of generality, we analyze time delays in row arrays $\Delta t_r$ and in column arrays $\Delta t_c$ from the same sound source. The incident angle of the time delays are with $\theta_r$ and $\theta_c$, respectively. Since the row direction and column arrays are vertical to each other, we can infer that $\theta_r + \theta_c = 90°$, which means $\sin^2\theta_r + \sin^2\theta_c = 1$. Then we can infer from Eq. (5):

$$\Delta t_r^2 + \Delta t_c^2 = \frac{f_s^2 d^2}{v^2}. \tag{9}$$

$f_s$, $d$ and $v$ are constants. According to the relationship between $\Delta t_r$ and $\Delta t_c$, we can match the peaks in different cross-correlation with each other.

*3) Beamforming Dimension Selection:* After matching the time delays of cross-correlation in different dimensions, we select a proper dimension to perform beamforming or assessment. If $\Delta t_r$ and $\Delta t_c$ are unique matches of each other, both two dimensions can be chosen as beamforming dimension. Another case is that *Peak Overlap* occurs, which means $\Delta t_r$ matches more than one time delays $\Delta t_{c1}, \cdots, \Delta t_{ck}$, or on the contrary. In this situation, one of the dimensions of the array cannot distinguish the overlapped peaks in cross-correlation. Since we need to assess the effect of beamforming with cross-correlation, we need to choose peak-distinguishable dimension for assessment. For example, Peak 1 ($\Delta t_r$ = -10 samples) of row arrays in Fig. 5(a) matches Peak 1 ($\Delta t_{c1}$ = 0 sample) and Peak 2 ($\Delta t_{c2}$ = 4 samples) of column arrays in Fig. 5(b) according to Eq. (9). In this situation, cross-correlation in column arrays separates the AoA of sound sources in a fine granularity. Thus, it is more suitable for assessment of beamforming effect.

In fact, there are multiple dimensions for the system to perform cross-correlation or assessment, including row, column, diagonal, etc. Cross-correlations can be performed in proper dimension with the 2D microphone array for better performance.

### *D. Beamforming*

Beamforming is performed according to the AoA inferred from corresponding time delay in the selected dimension. We first adjust the linear arrays in selected dimension to the direction of AoAs based on reference point. Then, we perform adaptive beamforming in the selected dimension and get one series of voice signal at the position of the reference point.

*1) Beamforming Direction Decision:* The first step of beamforming is to "steer" the linear microphone array to the input angles. For a linear microphone array, steering to a specific sound source $S$ emitting the signal $s(t)$ means adjusting the received signals $m_i(t)$ of the microphones to make the signals aligned according to the emitted signal $s(t)$. Suppose we steer row arrays to sound source $S$. As shown in Fig. 10, the incident angle is $\theta$ for row dimension and the frequency of the signal is $f$. Without loss of generality, we pay attention to the row array from $M_1$ to $M_4$. The middle point of the row is chosen as the steering center, which is denoted as $Y_1$. In far field propagation model, the time difference of arrival $\mathcal{T}_i$ between $M_i$ and the steering center $Y_1$ should be:

$$\mathcal{T}_i = (\frac{5}{2} - i) d \cos\theta / v. \tag{10}$$

Let the signal recorded by $M_i$ after steering denoted as $\overline{m}_i(t)$ (at $1', 2', 3', 4'$ in Fig. 10). $e^f_{m_i}$ represents the phase of $f$ frequency component of original signal $m_i(t)$ and $e^f_{\overline{m}_i}$ is the phase of $f$ frequency component of steered signal $\overline{m}_i(t)$. Then phase of $f$ frequency component $e^f_{\overline{m}_i}$ in steered signal $\overline{m}_i(t)$ should be:

$$e^f_{\overline{m}_i} = e^f_{m_i} \cdot e^{-j2\pi f(\frac{5}{2}-i)d\cos\theta/v}, \tag{11}$$

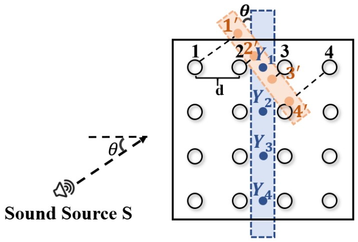


Fig. 10: Illustration of “Steering” in beamforming.

For broadband signals, we perform Fourier transform on signal $m_i(t)$ received by $M_i$ and get the phases $e^f$ in the form of complex numbers on different frequencies $f$. As voice signals are broadband signals, we can steer all frequency components of the signal $m_1(t)$ to the incident angle. In Fig. 10, $M_1$, $M_2$, $M_3$, $M_4$ are projected to $M'_1$, $M'_2$, $M'_3$, $M'_4$, respectively.

*2) Adaptive Beamforming:* After steering the linear microphone arrays, we perform adaptive beamforming in the selected dimension. Adaptive beamforming techniques use the characteristics of both the source and noise signals to suppress noise signals. Frost Beamformer[17] is an adaptive beamformer used to separate broadband signals. The algorithm applies an finite impulse response (FIR) filter to the signals of the microphones. Weights on the taps of the microphone array are adapted to minimize noise power in the array output. Suppose there are $N$ microphones and the FIR has $J \times N$ taps, $J$ taps for each microphone. Then the following output power needs to be minimized with the frequency constraints:

$$\begin{aligned} &\min_{\mathcal{W}} \quad E(\overline{m}(k)) = E[\mathcal{W}^T \overline{\mathcal{M}}(t)] \\ &subject\ to \quad \mathcal{F} = [f_1\ f_2\ ...\ f_J] = [1\ 0\ ...\ 0], \end{aligned} \tag{12}$$

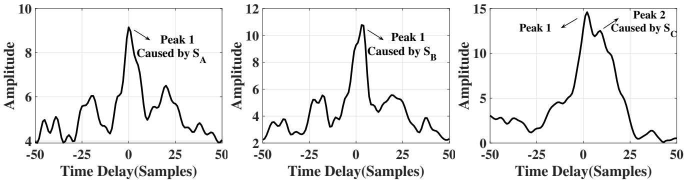
where $\mathcal{W}^T = [\omega_{11}, \omega_{12}, ..., \omega_{1J}, \omega_{21}, ..., \omega_{2J}, ..., \omega_{N1}, ..., \omega_{NJ}]$ are weights of the FIR, $\overline{\mathcal{M}}(t)^T = [\overline{m}_1(t), \overline{m}_1(t+\tau), ..., \overline{m}_1(t+(J-1)\tau), \overline{m}_2(t), ..., \overline{m}_2(t+(J-1)\tau), ..., \overline{m}_N(t), ..., \overline{m}_N(t+(J-1)\tau)]$ are received signals with different delays and $f_j = \sum_{n=1}^{N} \omega_{nj}$. The constraints $\mathcal{F}$ means there are no special frequency constraints for the final output. The weight matrix $\mathcal{W}$ can be inferred by Lagrange multipliers[41]. In our system, the number of microphones $N$ is 4 and we may change the number of the taps $J$ to achieve a balance between running time and performance.

Frost beamforming outputs one series of voice signals $m'(t)$ from steered signals and the equivalent position of the component from steering direction is the steering center. That means there exist a virtual microphone at $Y_1$ receiving the signal $m'_1(t)$ from sound source $S$. Frost beamforming is performed on 4 linear microphone arrays in selected dimension and we get four series of voice signals $m'_1, m'_2, m'_3, m'_4$ located at virtual microphones $Y_1, Y_2, Y_3, Y_4$.

### *E. BF-Correlation Assessment*

*1) BF-Correlation: BF-Correlation* performs cross-correlation between the outcomes of beamforming. The basic idea of adaptive beamforming is to suppress the signals out of interest, i.e., the voice signals out of the specified AoAs. Thus, the outcome of beamforming can be simplified as:

$$m'_l(t) = \sum_{i=1}^{n} \alpha_{i,l} s_i(t + \Delta t'_{i,l}).C \tag{13}$$



(a) Steering angle is 180°. The time delay of Peak 1 is 0 sample.

(b) Steering angle is 203°. The time delay of Peak 1 is 4 samples.

(c) Steering angle is 244°. The time Delay of Peak 2 is 9 samples.

Fig. 11: *BF-Correlation* for different steering angles.

$m'_l$ denotes the outcome of beamforming from the $l$th linear array. $\alpha_{i,l}$ is the gain for signal $s_i(t)$, and $\Delta t'_{i,l}$ is decided by steering. Suppose there are two series of beamformed signals $m'_p(t)$ and $m'_q(t)$, According to Eq. (3) and (6), we infer that

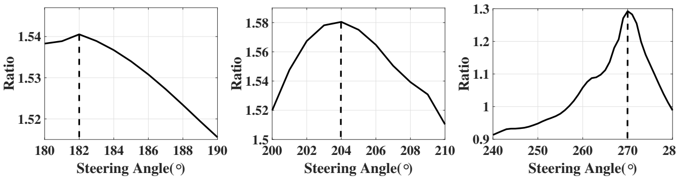
$$Cor_{m'_p(t), m'_q(t)} = \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_{i,p} \alpha_{k,q} Cor_{s_i(t+\Delta t'_{i,p}), s_k(t+\Delta t'_{k,q})}. \tag{14}$$

We find that the gain $\alpha_{i,p}$ and $\alpha_{i,q}$ of signal $s_i(t)$ are related to the amplitude of cross-correlation. If we perform beamforming towards $S_i$, the effect of beamforming can be assessed by the ratio $\alpha_{i,p}\alpha_{i,q} / \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_{i,p}\alpha_{k,q}$, which is equivalent to the ratio of *the amplitude of the corresponding peak to the average amplitude of all time delays* in *BF-Correlation*.

For example, we perform Frost beamforming with row arrays, $M_1, M_2, M_3, M_4$ and $M_{13}, M_{14}, M_{15}, M_{16}$, on three steering angle $180°, 203°$ and $244°$ corresponding to the coarse-grained AoAs in Fig. 5(b). After that, we perform *BF-Correlation* between the outputs of beamforming $m'_1$ and $m'_4$ in column dimension. Fig. 11 shows the *BF-Correlation* between $m'_1$ and $m'_4$ with different steering angles. We can see from the figures that the relative amplitude of the peaks are changed due to the beamforming in row dimension compared to Fig. 5(b). Peak 1 in Fig. 11(a), Peak 1 in Fig. 11(b), and Peak 2 in Fig. 11(c) correspond to Peak 1, Peak 2 and Peak 3 in Fig. 5(b), respectively. And the relative amplitude of the peaks in Fig. 11 are much higher than that in Fig. 5(b). Therefore, the ratio of the amplitude of the corresponding peak to the average amplitude of all time delays in BF-Correlation depicts the effect of beamforming.

*2) Local Search in Iteration Framework:* After calculating the ratio, we further perform beamforming on the steering angles around the initial coarse-grained AoA angle. Each steering angle corresponds to a ratio according to the *BF-Correlation* between outputs of beamforming. We can calculate ratios of different steering angles iteratively. Then, we use local search techniques to find the maximum ratio around the original steering angle. The searching regions and granularity can be modified according to the requirements of the precision. And the steering angle corresponding to the maximum ratio is the optimized AoA of the corresponding sound source.

For example, we calculate the ratios around the three AoAs estimated from cross-correlation. In Fig. 12, we draw figures between steering angles and ratios. In Fig. 12(a), the local maximum ratio appears when steering angle is 182°. The AoA of $S_A$ is then optimized from 180° to 182°. The other two AoAs can be optimized in the same way. Moreover, the



(a) Ratio for $S_A$ at different incident angles. (b) Ratio for $S_B$ at different incident angles. (c) Ratio for $S_C$ at different incident angles.

Fig. 12: Local Search for Optimum Solution.

voice separation have better performance in speech recognition system after the iteration of beamforming and *BF-Correlation*.

## V. Performance Evaluation

### A. Experimental Setup

We evaluate our system with the microphone array consisting of XMOS XU216 [42] microcontroller and 4 × 4 InvenSense ICS-41350 [43] microphones. The distance between any two adjacent microphones was 2.4 cm. We deployed the microphone array on the table in a meeting room, as shown in Fig. 1. By default, we deployed mobile phones as sound sources on tripods at the same height with the microphone array, while each sound source is kept 1.5∼2 m away from the microphone array with different intersection angles. The audios played by the sound sources are human voices. Each sentence recorded by the microphone array is regarded as one sample, and the average lasting time of the sentences is around 10 seconds. We recorded around 20 hours human voices for our experimental analysis.

**Metrics.** We evaluate the performance from both the estimation accuracy of sound direction and the recognition accuracy of the separated voices. For the direction estimation, we use the error of the AoA estimation compared with the groundtruth to evaluate. In the experiments, we set the search granularity of AoA to 0.1°. For the separated voices, we use iFLYTEK [44] Automatic Speech Recognition (ASR) systems to recognize them. For each separated voice, we record another comparison copy, when only the corresponding sound source is playing individually. Suppose $\gamma_s$ and $\gamma_c$ represents the recognition accuracy of the separated voice and the comparison copy based on iFLYTEK, then we use the $\gamma_s/\gamma_c$ to evaluate the relative accuracy, which is defined as *relative ASR ratios*.

### B. Macro Benchmark

We compared our system with Symphony [5], VoLoc [4] and cross-correlation methods. Symphony and VoLoc were performed with 4-mic linear array.

*Our solution achieves the best performance in AoA estimation and voice separation among the approaches.* Fig. 13(a) and 13(b) plot the average AoA errors and average relative ASR ratios of different approaches. The average AoA errors of our solution is 1.1° for single sound source and 1.5° for multiple sound sources, which are better than the existing methods. Here, VoLoc cannot work in multiple sound sources scenarios. The relative ASR ratios based on *BF-Correlation*, Symphony and cross-correlation are 88.2%, 74% and 50.4%, respectively. The reason for the outperforming of our method is that the other approaches analyze the signals in the time domain with large granularity and the received signals are interfered by *Peak Confusion*. Our approach can handle *Peak Confusion* and estimates the AoAs of the sound sources in a fine-grained manner. Fig. 13(c) and Fig. 13(d) plot the Cumulative Distribution Function (CDF) of AoA error and ratio of ASR accuracy of our solution. We find that 80% of the AoA estimation errors can be controlled less than 1.7° and only 20% of the accuracy ratios are below 74.8%. That means our approach can achieve good performance in most cases.

### C. Micro Benchmark

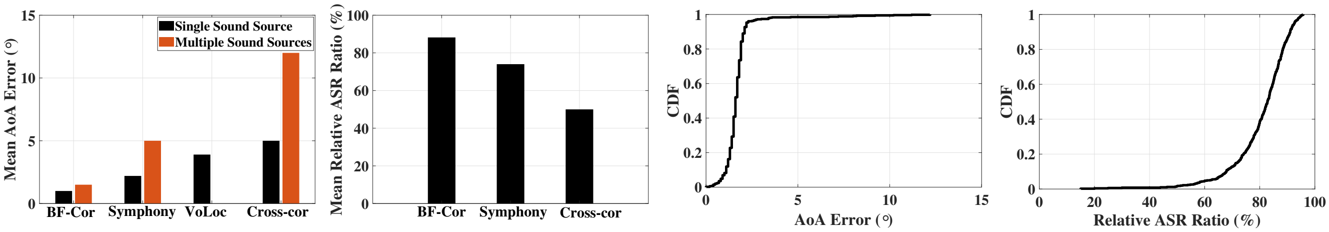
**1) Robustness to Different Numbers of Sound Sources.** *Our solution can achieve good performance when the number of sound sources in the scenario is less than 5.* To evaluate the robustness to the number of sound sources, we deploy 2, 3, 4 and 5 sound sources in the meeting room, respectively. Fig. 14(a) plots the AoA estimation errors and *relative ASR ratio*. We find that due to the strong interference among the voice signals, the larger number of sound sources is, the poorer the performance will be. When the number of sound source is 4, the average AoA error is 2.8° and the average relative ASR ratio is 76%. The performance of our solution gets unacceptable if there are 5 sound sources in the scenario.

**2) Robustness to Different Intersection Angles.** *Our system can efficiently separate the multiple sound sources, when the intersection angle between adjacent sound sources is over* 20°. We set two sound sources in the experiment and changed the intersection angles between the two sound sources from 10° to 90°. Fig. 14(b) shows that with the increase of the intersection angles, the performance of AoA estimation and voice separation gets better.

**3) Robustness to Different Noise Conditions.** *Our system can achieve high accuracy in AoA estimation and acceptable performance of voice separation in different noisy scenarios.* We conducted the experiments in three scenarios with different noise conditions, including quiet environment, noise from single direction and noise from all directions. In the quiet scenario, there are no sound sources except for the sound sources. We brought in the noise from air conditioner in the second scenario and brought in the noise of rains in the third scenario. Fig. 14(c) shows the impact of the noise. The noises mainly affect the performance of voice separation and slightly affect the performance of AoA estimation. Nevertheless, our system can still work with acceptable performance.

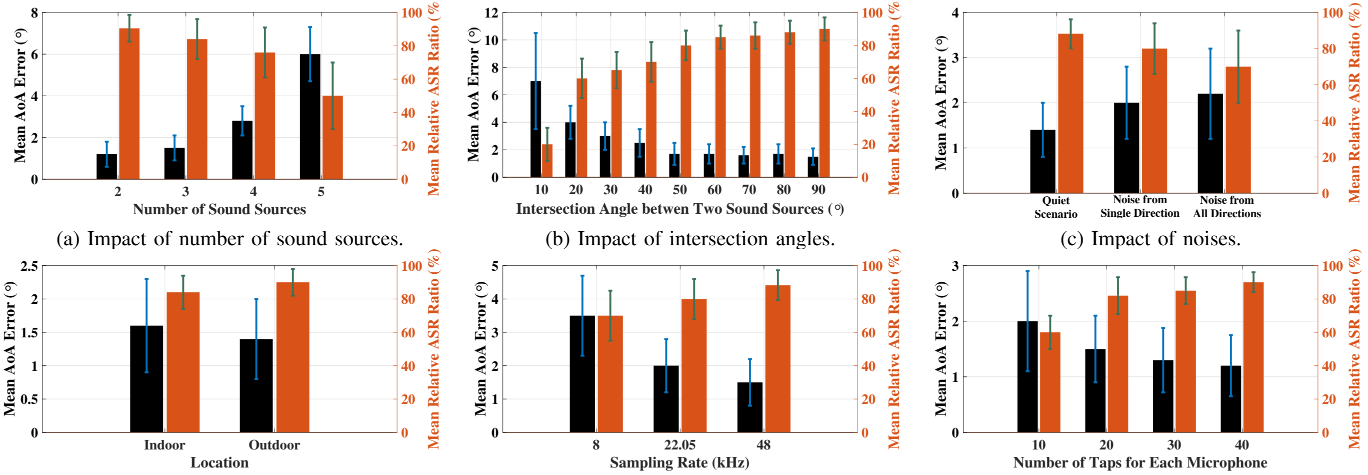
**4) Robustness to Different Multipath Conditions.** *Our system can achieve better performance outdoors because indoor environment has more complicated multipath effect.* The indoor experiments is performed in the meeting room. The main reflectors are walls, ceilings and tables. The outdoor experiments are performed in open with few reflectors except for the ground. The result is shown in Fig. 14(d). We infer that the multipath effect in the indoor environment is much more severe than the outdoor environment. Although spatial filters can suppress multipath effect, they cannot completely eliminate the multipath effect, which reduces the accuracy in the indoor environment.

**5) Robustness to Different Parameters.** *Higher sampling rate and larger number of taps can improve the performance*



(a) Comparisons of mean AoA errors. (b) Comparison of mean relative ASR ratios. (c) CDF of AoA Error. (d) CDF of relative ASR ratio.

Fig. 13: Macro benchmark.



(a) Impact of number of sound sources. (b) Impact of intersection angles. (c) Impact of noises.

(d) Impact of locations. (e) Impact of sampling rate. (f) Impact of number of taps.

Fig. 14: Micro benchmark.

*of voice separation.* We further changed the sampling rate and the tap number in our system, to evaluate the impact of the software parameters. Fig 14(e) and 14(f) show the performance of the system with different sampling rates and tap numbers. With the increase of sampling rate and number of taps, the performance becomes better and the computational consumption increases. Particularly, the number of taps mainly affect the voice separation, while the sampling rate affect both the AoA estimation and voice separation.

## VI. Discussion

**Multipath Interference.** The multi-path effect in indoor environment could create "virtual" sound sources, i.e., multiple sound sources emit the same voice signals from different directions with different time delay. This could further interfere with the AoA estimation and subsequent voice separation. To tackle this issue, we apply spatial filters for voice separation, which retain the signals from the steering direction and suppress the signals from other directions. Thus, in BF-correlation analysis, the voice signals from the original sound source cannot be influenced by the virtual sound source in other directions.

**Computational Cost of BF-Correlation.** The computational cost of BF-Correlation is high since iterative cross-correlation and beamforming are performed in the algorithm. To tackle this issue, we split signals into smaller segments. BF-Correlation is post-processing of voice signals, the system can first record all the voice signals and process afterwards. We can reduce the computational cost of BF-Correlation by segmenting the signals with suitable length. Meanwhile, lower sampling rate and matrix calculation optimization may reduce the computational cost as well.

## VII. Conclusion

In this paper, we propose a novel 2D microphone array-based system to perform voice separation in multiple sound source scenarios. A BeamForming-based cross-Correlation (*BF-Correlation*) scheme is presented for accurate AoA estimation and adaptive beamforming. We analyze the *peak confusion* issues for *cross correlation* via thorough empirical study and modeling, and propose to use *BF-Correlation* to accurately estimate the AoA of multiple sound sources and automatically optimize voice separation in our iterative framework. We implemented a prototype system and evaluate the performance in real environments. Experiments show that the average AoA error is 1.4° and the average ratio of automatic speech recognition accuracy is 90.2% in the presence of three sound sources.

## Acknowledgment

This work is supported partially by National Natural Science Foundation of China under Grant Nos. 61832008, 61872174, 61902175, 61872173; JiangSu Natural Science Foundation under Grant Nos. BK20190293 and the Key K&D Program of Jiangsu Province under Grant BE2020001-3. This work is partially supported by Collaborative Innovation Center of Novel Software Technology and Industrialization. This work is also supported by Australian Research Council (ARC) Discovery Project grant DP190101888. Lei Xie is the corresponding author.

# References

[1] Wikipedia. (2021) Amazon Echo. [Online]. Available: https://en.wikipedia.org/wiki/Amazon_Echo

[2] Wikipedia. (2021) Apple Homepod. [Online]. Available: https://en.wikipedia.org/wiki/HomePod

[3] Wikipedia. (2021) Google Home. [Online]. Available: https://en.wikipedia.org/wiki/Google_Nest_(smart_speakers)

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

[5] W. Wang, J. Li, Y. He, and Y. Liu, "Symphony: localizing multiple acoustic sources with a single microphone array," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 82–94.

[6] G. Wang, C. Qian, L. Shangguan, H. Ding, J. Han, N. Yang, W. Xi, and J. Zhao, "Hmrl: Relative localization of rfid tags with static devices," in *2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2017, pp. 1–9.

[7] C. Chao, 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, pp. 55:1–24, 06 2021.

[8] Z. An, Q. Lin, L. Yang, and Y. Guo, "Revitalizing ultrasonic positioning systems for ultrasound-incapable smart devices," *IEEE Transactions on Mobile Computing*, vol. 20, no. 5, pp. 2007–2024, 2020.

[9] Q. Lin, Z. An, and L. Yang, "Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices," in *The 25th Annual International Conference on Mobile Computing and Networking*, 2019, pp. 1–16.

[10] H. Zhu, Y. Zhang, Z. Liu, X. Wang, S. Chang, and Y. Chen, "Localizing acoustic objects on a single phone," *IEEE/ACM Transactions on Networking*, 2021.

[11] H. Ding, J. Han, C. Qian, F. Xiao, G. Wang, N. Yang, W. Xi, and J. Xiao, "Trio: Utilizing tag interference for refined localization of passive rfid," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 828–836.

[12] C. Cai, H. Pu, P. Wang, Z. Chen, and J. Luo, "We Hear Your PACE: Passive Acoustic Localization of Multiple Walking Persons," in *Proc. of ACM Ubicomp*, 2021, pp. 55:1–24.

[13] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[14] I. Rec, "P. 800.1, mean opinion score (mos) terminology," *International Telecommunication Union, Geneva*, 2006.

[15] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[16] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the em algorithm," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 4, pp. 477–489, 1988.

[17] O. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[18] H. Pu, C. Cai, M. Hu, T. Deng, R. Zheng, and J. Luo, "Towards Robust Multiple Blind Source Localization Using Source Separation and Beamforming," *Sensors*, vol. 21, no. 2, pp. 1–10, 2021.

[19] D. Ciuonzo, G. Romano, and R. Solimene, "Performance analysis of time-reversal music," *IEEE Transactions on Signal Processing*, vol. 63, no. 10, pp. 2650–2662, 2015.

[20] E. Fishler and H. V. Poor, "Estimation of the number of sources in unbalanced arrays via information theoretic criteria," *IEEE Transactions on Signal Processing*, vol. 53, no. 9, pp. 3543–3553, 2005.

[21] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[22] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 530–541, 2020.

[23] 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," *arXiv preprint arXiv:1804.03619*, 2018.

[24] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 667–673.

[25] K. Tan, Y. Xu, S.-X. Zhang, M. Yu, and D. Yu, "Audio-visual speech separation and dereverberation with a two-stage multimodal network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 542–553, 2020.

[26] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of machine learning research*, vol. 3, no. Jul, pp. 1–48, 2002.

[27] H. Nguyen and R. Zheng, "Binary independent component analysis with or mixtures," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3168–3181, 2011.

[28] A. Painsky, S. Rosset, and M. Feder, "Generalized binary independent component analysis," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 1326–1330.

[29] X. Ji, M. Yu, J. Chen, J. Zheng, D. Su, and D. Yu, "Integration of multi-look beamformers for multi-channel keyword spotting," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7464–7468.

[30] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[31] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 6, pp. 1391–1400, 1986.

[32] L. Wang and A. Cavallaro, "Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles," *IEEE Sensors Journal*, vol. 17, no. 8, pp. 2447–2455, 2017.

[33] R. Ma, G. Liu, Q. Hao, and C. Wang, "Smart microphone array design for speech enhancement in financial vr and ar," in *2017 IEEE SENSORS*. IEEE, 2017, pp. 1–3.

[34] X. Zhang, Z.-Q. Wang, and D. Wang, "A speech enhancement algorithm by iterating single-and multi-microphone processing and its application to robust asr," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 276–280.

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

[36] P. Poozesh, K. Aizawa, C. Niezrecki, J. Baqersad, M. Inalpolat, and G. Heilmann, "Structural health monitoring of wind turbine blades using acoustic microphone array," *Structural Health Monitoring*, vol. 16, no. 4, pp. 471–485, 2017.

[37] C. Cai, Z. Chen, H. Pu, L. Ye, M. Hu, and J. Luo, "Acute: acoustic thermometer empowered by a single smartphone," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 28–41.

[38] C. Cai, H. Pu, L. Ye, H. Jiang, and J. Luo, "Active acoustic sensing for hearing temperature under acoustic interference," *IEEE Transactions on Mobile Computing*, 2021.

[39] K. Sun, T. Zhao, W. Wang, and L. Xie, "Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 591–605.

[40] Wikipedia. (2021) Cross-correlation. [Online]. Available: https://en.wikipedia.org/wiki/Cross-correlation

[41] A. E. Bryson and Y.-C. Ho, *Applied optimal control: optimization, estimation, and control*. Routledge, 2018.

[42] XMOS. (2018) Xmos xu216-512-tq128. [Online]. Available: https://www.xmos.ai/file/xu216-512-tq128-datasheet/

[43] Invensense. (2021) Invensense ics-41350. [Online]. Available: https://product.tdk.com/en/

[44] iFLYTEK. (2021) iflytek automatic speech recognition. [Online]. Available: https://global.xfyun.cn/