M^{2}SILENT: Enabling Multi-user Silent Speech Interactions via Multi-directional Speakers in Shared Spaces

Juntao Zhou
Department of Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, China
juntaozhou@sjtu.edu.cn

Dian Ding*
Department of Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, China
dingdian94@sjtu.edu.cn

Yijie Li
School of Computing
National University of Singapore
Singapore, Singapore
yijieli@nus.edu.sg

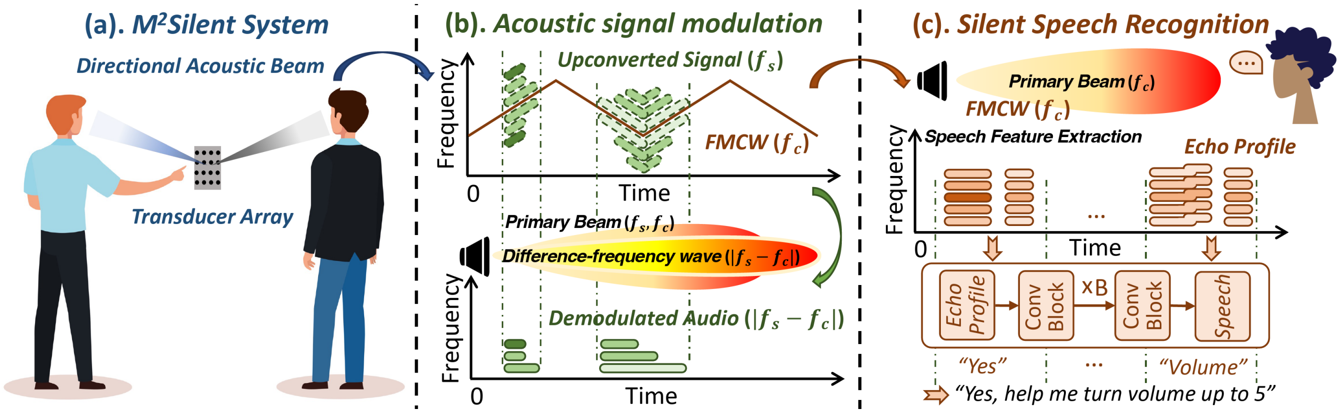
Yu Lu
Department of Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, China
yulu01@sjtu.edu.cn

Yida Wang
Shanghai Jiao Tong University
Shanghai, China
yidawang@sjtu.edu.cn

Yongzhao Zhang
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
zhangyongzhao@uestc.edu.cn

Yi-Chao Chen*
Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, China
yichao0319@gmail.com

Guangtao Xue
Department of Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, China
gt_xue@sjtu.edu.cn



(a). M²Silent System
Directional Acoustic Beam
Transducer Array
(b). Acoustic signal modulation
Upconverted Signal (f_s)
Frequency
0
Time
FMCW (f_c)
Primary Beam (f_s, f_c)
Difference-frequency wave ($|f_s - f_c|$)
Demodulated Audio ($|f_s - f_c|$)
Frequency
0
Time
(c). Silent Speech Recognition
Primary Beam (f_c)
FMCW (f_c)
Speech Feature Extraction
Echo Profile
Frequency
...
0
Time
Echo Profile
Conv Block
xB
Conv Block
Speech
"Yes"
...
"Volume"
"Yes, help me turn volume up to 5"

Figure 1: M^{2}Silent includes the multi-directional speaker system, acoustic signal modulation, and silent speech recognition. The illustration shows how FMCW is used as a carrier for simultaneous audio transmission and silent speech sensing.

Abstract

We introduce M^{2}Silent, which enables multi-user silent speech interactions in shared spaces using multi-directional speakers. Ensuring privacy during interactions with voice-controlled systems presents significant challenges, particularly in environments with multiple individuals, such as libraries, offices, or vehicles. M^{2}Silent addresses this by allowing users to communicate silently, without producing audible speech, using acoustic sensing integrated into directional speakers. We leverage FMCW signals as audio carriers, simultaneously playing audio and sensing the user's silent speech.

*Both Dian Ding and Yi-Chao Chen are corresponding authors.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714174

To handle the challenge of multiple users interacting simultaneously, we propose time-shifted FMCW signals and blind source separation algorithms, which help isolate and accurately recognize the speech features of each user. We also present a deep-learning model for real-time silent speech recognition. M^{2}Silent achieves Word Error Rate (WER) of 6.5% and Sequence Error Rate (SER) of 12.8% in multi-user silent speech recognition while maintaining high audio quality, offering a novel solution for privacy-preserving, multi-user silent interactions in shared spaces.

CCS Concepts

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

Keywords

Silent speech interaction, Multi-directional speaker, Air nonlinearity, Acoustic sensing

ACM Reference Format:
Juntao Zhou, Dian Ding, Yijie Li, Yu Lu, Yida Wang, Yongzhao Zhang, Yi-Chao Chen, and Guangtao Xue. 2025. M^{2}SILENT: Enabling Multi-user Silent Speech Interactions via Multi-directional Speakers in Shared Spaces. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3706598.3714174

1 Introduction

As voice control systems rapidly integrate into our daily lives, maintaining privacy during interactions has become increasingly important. While voice assistants on smartphones and smart devices enhance user convenience, they often face limitations in settings such as vehicles, museums, and offices due to privacy concerns, discomfort from speaking openly, and challenges in noisy environments. Silent speech interfaces (SSI) [20, 59, 99] address these issues by enabling users to communicate without producing audible speech, making them useful for speech impairments or silent communication needs. Many SSIs rely on wireless signals to detect speech movements, with current research exploring electromyography (EMG) [47], ultrasound imaging [102], and video-based lip reading [87]. These techniques hold promise for improving privacy and facilitating more discreet user interactions.

In real-world scenarios such as museums or driving, SSIs must not only detect silent speech but also deliver feedback (e.g., exhibit descriptions or navigation instructions). Nevertheless, SSIs that rely on camera or millimeter-wave [94] require supplementary devices to facilitate two-way communication with users. In contrast, acoustic-based systems can employ frequency division multiplexing, using low frequencies for audio transmission and high frequencies (around 20 kHz) for motion detection, thus integrating silent speech recognition with audio output. However, the spherical wave propagation in these acoustic systems [85] can create noise in public spaces, raising privacy issues. Low-frequency speaker arrays [69] offer directional sound but are often bulky and still suffer from leakage. Moreover, multiple users typically require interaction in public settings, making individual systems per user impractical. Current acoustic SSIs focus on a single user, such as those integrated with smartphones [74], smartwatches [97], or smart glasses [101], and necessitate close proximity to the device. While these systems offer personalization and enhanced privacy, they are incapable of serving individuals not equipped with such devices, such as museum visitors or large crowds in public spaces. This limitation renders them unsuitable for open and dynamic environments.

To this end, we propose M^{2}Silent, a novel acoustic platform for multi-user simultaneous "private" audible signal transmission and "concealed" acoustic silent speech recognition. Parametric arrays [95] offer a promising approach for directed sound transmission. This method modulates low-frequency sound waves onto high-frequency carriers, where the low-frequency audio is demodulated through air nonlinearity and maintains the high directivity of the high-frequency sound waves. For instance, MuDiS [45] is capable of delivering audible sound to users from multiple directions without leakage, ensuring silence in other areas. However, it supports only one-way communication and lacks sensing capabilities. As shown in Fig. 1, M^{2}Silent delivers focused audio to multiple users while simultaneously capturing their silent speech. Fig. 2 illustrates various applications. Fig. 2(a) shows an in-car space where M^{2}Silent interact simultaneously with the driver and passenger. The driver engages in silent interaction related to navigation, while the passenger interacts with music without interference, maintaining a quiet environment inside the car. In a busy museum (Fig.2(b)), visitors can receive personalized audio feedback on exhibits without disturbing others. At a bank counter (Fig. 2(c)), users can silently convey sensitive transaction information and hear private responses. By eliminating the need for wearable or supplemental devices, M^{2}Silent reduces interaction costs and enhances privacy in shared spaces.

However, implementing such a multi-user silent speech interaction system in a shared space poses several challenges. First, traditional acoustic systems use frequency division multiplexing to transmit both sensing signals (around 20kHz) and low-frequency sound waves simultaneously. However, since parametric array speakers can only operate within a narrow ultrasonic frequency band, embedding the sensing signals for silent speech recognition without interfering with the original directional playback function of the multi-directional speakers is a challenge. Second, in multi-user scenarios, the system needs to support simultaneous interaction from multiple users. However, the signals received by the same microphone may have overlapping lip movement features from different users. Furthermore, it is necessary to ensure both silent speech recognition functionality and real-time system performance in real-world use.

This paper aims to implement multi-user silent speech interaction in a shared space using a multi-directional speaker. To detect users' silent speech while simultaneously playing audio signals, we innovatively use Frequency Modulated Continuous Wave (FMCW) signals as the audio carrier. After undergoing nonlinear demodulation in the air, the transmitted signal provides clear audio to the user. Meanwhile, the FMCW signal as the carrier is reflected and captures silent speech from multiple individuals. To separate different users' silent speech features, we transmit FMCW signals with time offsets in different directions, which results in unique features for each user appearing on different spectrums. To further address the potential feature overlap, we introduce a blind source separation algorithm to cleanly isolate the features of each user. Finally, we

"to the Plaza"
"volume down"
"artist info"
"where is hall B"
"voice guide"
$$ "transaction to..."

(a) **In-car space:** M^{2}Silent allows the driver to engage in silent interaction with the navigation system while the passenger controls the music without interference.

(b) **Museum:** Visitors can interact with M^{2}Silent to ask for exhibit information or directions, while other visitors are not disturbed by the interactions.

(c) **Bank counter:** A user communicates sensitive information of transactions through M^{2}Silent, allowing private conversations between the user and the teller.

Figure 2: Potential use cases of M^{2}Silent.

utilize a deep residual model, SilentMatch, to accurately recognize users' silent speech. Real-time interaction is facilitated through the continuous processing of input sequences using a sliding window approach, ensuring seamless and efficient recognition.

In summary, the main contributions of this paper are as follows:

- To the best of our knowledge, M^{2}Silent appears to be the first silent speech interaction system for open environments, using multi-directional speakers to enable device-free non-intrusive multi-user interaction. The system suits quiet and private settings and pushes SSI applications toward more public use cases.
- We propose a synchronous modulation technique leveraging air nonlinearity, which innovatively employs frequency-modulated continuous wave (FMCW) as audio carriers. This approach enables directional loudspeakers to simultaneously transmit audio and sensing signals, facilitating silent speech recognition for multiple users.
- We propose time-shifted FMCW on directional acoustic beams for different users, utilizing a blind source separation algorithm for simultaneous multi-user interactions. Additionally, we implement a silent word recognition model to extract lip movement features, employing a sliding window approach to facilitate sentence-level speech recognition.
- The extensive experiments in real-world environments demonstrate that M^{2}Silent achieves a low word error rate (WER) of 6.5% and a sequence error rate (SER) of 12.8% in multi-user silent speech recognition while maintaining high audio quality, as reflected by a PESQ score of 2.81.

2 Related Work

2.1 Silent Speech Recognition

2.1.1 Acoustic-based Methods. Acoustic-based methods for silent speech recognition have gained considerable attention for their ability to capture subtle speech-related movements. For example, SoundLip [98] uses acoustic sensing for silent lip interaction, recognizing both individual words and continuous sentences. EarCommand [34] leverages ear canal deformations for silent speech detection, illustrating the feasibility of everyday wearable integration. EchoSpeech [101] highlights non-intrusiveness by employing minimally obtrusive eyewear for discrete and continuous speech recognition. HPSpeech [99] relies on commodity headphones to sense jaw movements, indicating the versatility of acoustic sensing across diverse form factors. Meanwhile, Lipwatch [97] and EarSSR [77] continue to advance the field with smartwatch- and earphone-based silent speech recognition, emphasizing user convenience and seamless integration into common wearable technologies.

In contrast to systems that rely on additional wearable devices such as headphones [34, 77, 99], smartwatches [97], or glasses [101] which users may find uncomfortable or aesthetically displeasing [32, 91], M^{2}Silent facilitates silent voice interaction without requiring users to wear any external sensors. Other approaches use smartphones [98], requiring the user's lips to be very close to the phone, which makes them impractical in scenarios where the user's hands are occupied (e.g., holding an umbrella or writing). In contrast, M^{2}Silent only requires the user to face a multi-directional speaker from a distance for silent speech interactions without additional effort. Moreover, existing systems are highly personalized and not suitable for simultaneous use by multiple users in public environments. However, it is essential to provide voice interaction for numerous users in various scenarios, such as museums or vehicles.

2.1.2 Beyond Acoustic Methods. Other silent speech recognition methods employ diverse sensing techniques. For example, mSilent [94] leverages mmWave radar with deep learning for fine-grained speech features in various conversational contexts, while Lee et al. [43] and TWLip [107] use IR-UWB and coherent SISO radar, respectively, to enable contactless silent speech recognition. Camera-based approaches also feature prominently. SpeeChin [100] uses an IR camera on a smart necklace to capture neck and face images for silent speech commands, and LipLearner [74] employs contrastive and few-shot transfer learning with mobile device cameras to facilitate customizable recognition. MELDER [58] further emphasizes real-time processing and optimization on mobile devices, achieving high accuracy and speed. Beyond these methods, IMUs and magnetic sensing offer additional alternatives. Srivastava et al. [71] recognizes unvoiced commands via a twin-IMU wearable that tracks jaw motion, and Hofe et al. [24] employs magnetic sensors for a small-vocabulary silent speech interface tailored to users with speech impairments. Meanwhile, sEMG-based systems [47] capture muscle activity to decode silently mouthed phrases, demonstrating the technology's potential in handling more complex vocabularies.

Although these non-acoustic sensing methods, such as millimeter-wave [94], IR-UWB radar [43], coherent radar [107], are capable of supporting longer ranges, they cannot facilitate bidirectional acoustic communication, meaning users cannot receive audible feedback. In contrast, M^{2}Silent is a highly integrated acoustic device that allows users to emit silent speech and hear acoustic feedback simultaneously without investing in additional sensing equipment. On the other hand, camera-based solutions on smartphones [58, 74], while enabling bidirectional interaction, have raised ongoing privacy concerns. They are also susceptible to lighting conditions, and these devices are private and meant for individual use only. Other sensing methods, such as IR camera [100], IMUs [71], magnetic fields [24], or sEMG [47], require users to wear the device around their necks or extremely close to their faces or mouths, which can be uncomfortable. In contrast, M^{2}Silent will not cause discomfort to users.

2.1.3 Difference from Whispered Speech Recognition. Various studies focus on whispered speech recognition [13, 22, 64, 65]. The key distinction between whispered speech and silent speech lies in the fact that whispered speech is audible but at a lower volume and requires the user to be very close to the microphone. In contrast, silent speech is completely inaudible, relying on non-acoustic signals such as lip movements, muscle activity, or skin vibrations, which require specialized sensors like EMG [47] or speaker microphones [98]. WESPER [65] suggests that whispered speech, being directly captured by microphones, has the potential to be converted into normal speech with reduced model training costs. However, whispered speech's low volume makes it difficult to capture in noisy environments or from a distance. Silent speech, on the other hand, offers the advantage of usability in noisy environments, with systems like M^{2}Silent employing multi-directional speakers to enable long-distance silent speech recognition even in open scenarios.

2.2 Directional Speaker

2.2.1 Principles of Directional Speakers. Directional speakers, particularly those using air nonlinearity, have been extensively studied in acoustic engineering. Westervelt [89] and Yang [92] demonstrated that nonlinear propagation in air enables ultrasound demodulation, resulting in highly focused sound beams. Early projects by Woodynorris and Yoneyama [53, 93] led to practical devices such as the SoundLazer [37], employing air nonlinearity to deliver tightly directed audio. Originally developed for underwater applications like sub-bottom profiling [26] and communication [39], parametric arrays were later adapted for air-based usage. This approach provides precise, compact sound projection in directional speakers [84, 95], as well as in sound spot generation [54, 105], targeted communication [4, 11], and personalized sound fields [63, 106].

2.2.2 The Implementation of Multi-directional Speakers. Multi-directional speakers employ phased arrays via space or time division multiplexing. Early work [69] relied on large low-frequency devices for multi-angle sound projection, constrained by significant size requirements. Ultrasonic systems [10, 67] use air nonlinearity and multi-beamforming but typically emit the same audio in multiple directions and allow only a limited number of angles. MuDiS [45] addresses these issues by introducing a specialized ultrasonic transducer cell structure to expand steering angles while minimizing leakage, thereby achieving wide-angle digital steering and more flexible multi-directional capabilities.

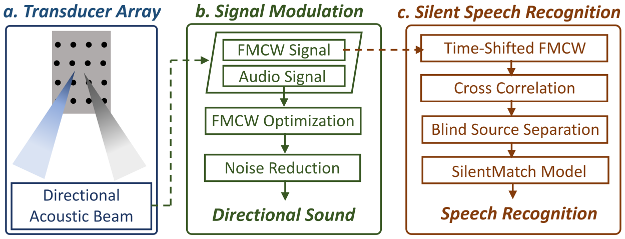


Figure 3: System overview. M^{2}Silent consists of three core components: (a) a transducer array producing multi-directional beams for focused sound transmission, (b) a signal modulation stage that modulates audio onto FMCW signals with optimization and noise reduction, and (c) a silent speech recognition module employing time-shifted FMCW, cross-correlation, blind source separation, and the SilentMatch model to accurately recognize multiple users' silent speech.

3 M^{2}Silent Framework

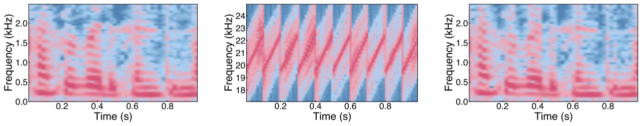
3.1 System Overview

In this paper, we introduce M^{2}Silent (Fig. 3), a system that enables silent speech interactions for multiple users in shared spaces by leveraging a multi-directional speaker [45] (see Sec.A.1 for more implementation details). M^{2}Silent employs Frequency Modulated Continuous Wave (FMCW) signals [73] both as an audio carrier and a sensing mechanism, simultaneously broadcasting audio while detecting silent speech via reflections from users' lip movements and facial dynamics. To support multiple users, M^{2}Silent introduces time-shifted FMCW signals in various directions, enabling separation and recognition of silent speech from different directions. A blind source separation algorithm further isolates individual speech signals. Additionally, the system incorporates a silent word recognition model with a sliding window for real-time lip movement analysis. By combining directional speakers, FMCW signals, and deep learning, M^{2}Silent ensures high-quality audio transmission and accurate silent speech recognition for multi-user interactions.

3.2 Empowering Directional Speakers with Sensing

In this section, we introduce how to enable directional speakers with sensing capabilities using FMCW signals.

3.2.1 Using FMCW Signal as Audio Carrier. Considering an FMCW signal that consists of multiple chirps within one period, its time-domain expression can be written as:



(a) Modulated audio signal. (b) Signal after modulation. (c) Demodulated signal.

Figure 4: (a) A 1-second audio clip of a female voice is shown, representing the original audio used for modulation. (b) The low-frequency signal has been modulated onto the FMCW signal. (c) The signal received after air nonlinearity demodulation is shown, closely resembling the original signal.

$$FMCW(t) = \sum_{n=0}^{N-1} \text{rect}\left(\frac{t - nT_c}{T_c}\right) \cdot \cos\left[2\pi\left(f_0 + \frac{B}{2T_c}(t - nT_c)\right)(t - nT_c)\right]$$

where N is the number of chirps, T_c is the chirp duration, f_0 is the starting frequency, B is the bandwidth, and n is the chirp index. To use FMCW as a carrier in directional speakers, the audio content $m(t)$ is modulated as:

$$s(t) = [1 + m(t)]\, FMCW(t)$$

With a center frequency of $21kHz$, a bandwidth of $3kHz$, and a chirp length of $0.1s$, the modulated audio (e.g., a voice clip shown in Fig. 4(a)) demonstrated successful low-frequency modulation onto the FMCW signal (Fig. 4(b)). After demodulation, the received signal closely resembled the original, as shown in Fig. 4(c).

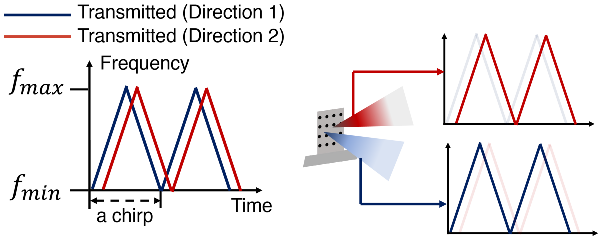
However, the sound quality significantly degraded, with a perceptual evaluation of speech quality (PESQ, a metric to assess speech quality mentioned in Sec. 5.1.2) score dropping by about 1. This degradation, attributed to the rapid frequency changes in FMCW, causes instability. Therefore, we selected the FMCW signal with the best auditory sense by verifying FMCW signals with different waveforms, chirp lengths, and bandwidths (see Sec. A.2 for details). We empirically choosed an FMCW signal with a linear triangle shape, a bandwidth of $2kHz$, and a chirp length of $0.25s$ as the carrier signal and used it for sensing simultaneously.

3.2.2 Optimization-based Noise Reduction in Demodulation. While our designed FMCW signal improves audio quality, its time-varying frequency still induces variations in perceived sound intensity due to the ultrasonic array's frequency response. Moreover, nonlinear distortions from ultrasonic modulation must be eliminated. To address these issues, we introduce an optimization method that fine-tunes the source audio before modulation.

Suppose the original audio signal in the frequency domain is $x(f)$, where f is the frequency. Our fine-tuning is multiplying the frequency domain of the audio by optimizable amplitude and phase coefficients, $A(f)$ and $\phi(f)$, yielding: $\hat{x}(f) = A(f) \cdot x(f) \cdot e^{j\phi(f)}$. Our goal is to make the demodulated audio closely match the original signal after nonlinear distortion. We formulate the optimization objective as:

$$\min \|\mathcal{LP}\left\{[(1 + \hat{x})FMCW(t)]^2\right\} - x\|_2$$

where $\mathcal{LP}\{\cdot\}$ is a function that models the audible portion of the signal after accounting for the speaker's frequency response and subsequent nonlinear demodulation. We use gradient descent for optimization. Optimizing one second of audio requires only $0.032s$, and with streaming, this meets real-time playback requirements.



(a) Time-shifted FMCW signals. (b) Multiple FMCW signals transmitted in different directions.

Figure 5: Time-shifted FMCW signal emission in multiple directions.

3.3 Multi-User Silent Speech Feature Extraction and Segmentation

In the previous section, we used FMCW signals to maintain the speaker's directional capabilities. Next, we utilize these signals for interaction, capturing real-time facial and lip dynamics of multiple users to enable multi-user silent speech recognition.

3.3.1 Acoustic-based Silent Speech Feature. In this study, silent speech, characterized by inaudible articulatory movements, is captured using FMCW signals. M^2SILENT extracts features by cross-correlating transmitted and received signals to produce echo frames [86, 101]. Through differential processing, these frames reveal subtle facial and muscle dynamics essential for decoding speech (the algorithm for silent speech extraction is detailed in Sec. A.3).

However, when dealing with **multiple users**, the signal spectrum becomes mixed, causing the silent speech features of each individual to overlap. The similar strength of these mixed signals makes it difficult to separate them using traditional methods. Thus, a method is required to effectively isolate these features.

3.3.2 Time-shifted FMCW Signal Emission in Multiple Directions. When multiple users are speaking silently, the features will be aliased together. Can we separate the features with only **one single microphone** by changing the way the FMCW signals are emitted in different directions? We propose a segmentation method in which the FMCW carriers transmitted in different directions have a certain time offset (Fig. 5(a)) so that the features of different users can be separated directly based on the separated cross-correlation peaks.

We assume N users interact with M^2SILENT in N directions (Fig. 5(b) shows the two-user case). In the signal sent to each user, we add a cyclic time shift to the FMCW signal. The time shift for the i-th user is $(i-1)t_{\text{shift}}$. The silent speech feature of the i-th user will be carried by the FMCW signal after $(i-1)t_{\text{shift}}$. At the receiver, all reflected signals are captured by a single microphone. We cross-correlate the reflected signal with the FMCW signal sent to the first user without any time shift. The sampling interval between the silent speech features of the $(i+1)$-th user and the i-th user in the

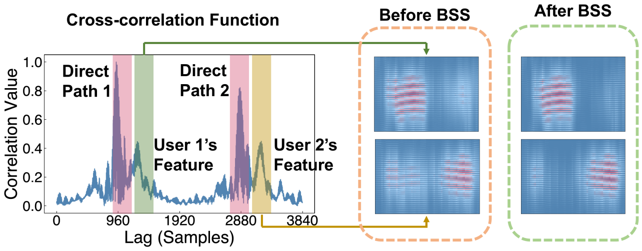


Figure 6: Blind source separation (BSS) [14] applied to separate features of multiple users. The diagram shows the echo profiles before and after applying BSS to differentiate between users' silent speech features.

cross-correlation function is

$$\Delta\text{Samples} = sr \cdot t_{\text{shift}},$$

where sr is the sampling rate, $96kHz$. Fig. 6 shows the extracted cross-correlation function in the case of two users. Since we use a time shift of $0.02s$, the silent speech features of the second user are shifted by 1920 sampling points compared to the first user. By extracting the sample points corresponding to each user, we can obtain the silent speech features of each user.

Nevertheless, we still observe slight leakage of each user's features in others' echo profiles. This occurs because the FMCW carrier, which is much stronger than the audio component, leaks in both directions and mixes the reflected signals. To tackle this, we propose a blind source separation method to decompose each user's silent speech features and enhance recognition performance.

3.3.3 Blind Source Separation of Mixed Features. As shown in Fig. 6, we use the blind source separation (BSS) method to segment the mixed features. We first analyze the characteristics of the mixed features to illustrate the applicability of the BSS method and then explain how we use this method for feature segmentation.

Assuming that the multi-directional speaker serves users in N different directions, for the i-th user, its silent speech feature F_i^{mix} on the spectrum graph can be expressed as

$$F_i^{mix} = \omega_i F_i + \sum_{j \neq i}^{N} \widetilde{\omega}_{ij} F_j$$

where F_i is the clean feature of the i-th user, ω_i is the amplitude weight corresponding to the main lobe, and $\widetilde{\omega}_{ij}$ is the amplitude weight corresponding to the leakage of the other j-th user in the i-th direction, which is different for each other user because the intensity of the side lobe changes with the angle in the beam pattern. For all users, we formulate this in matrix form as $\mathbf{F}^{mix} = A\mathbf{F}$, where A is the mixing matrix representing how the sources combine into the observed signals, $\mathbf{F}^{mix}$ is the matrix of all users' F^{mix} combined.

We apply the FastICA algorithm [42] to perform Blind Source Separation to recover the original source signals from the observed mixed signals, and finally get the silent speech features for all users. A detailed description of the blind source separation algorithm can be found in Sec. A.4.

3.4 Streaming Silent Speech Recognition

M^{2}Silent is designed for real-time silent speech processing with deep learning, enabling rapid responses. We treat streaming inputs as word sequences and adopt a compact word recognition model [46] for quick training and transfer learning, reducing complexity and enhancing portability. For streaming, we define time windows based on natural speaking rates [25], sliding over silent speech features with minimal overlap.

3.4.1 Word Recognition Model. We propose SilentMatch for silent word recognition combined with streaming processing to enable real-time silent speech recognition.

Word feature extraction. Based on Sec. 3.3, we further refine the feature extraction approach by focusing on the processing of silent word features. Specifically, in the time domain, we capture continuous features over short periods by selecting a time window with the same time as every chirp, with a $0.1s$ step size for sliding window feature extraction. Based on average human speech speed, we empirically set the length of each possible word to $1s$. Additionally, in the frequency domain, we use Fbank to map the data into a 64-dimensional frequency space, compressing the model's frequency domain features to avoid redundancy from overly similar nearby frequency features.

Word recognition. The extracted features are then fed into SilentMatch model, which is inspired by the framework presented in [46]. This network is essentially a convolutional neural network, a structure that has been employed in previous silent speech recognition studies [97, 101]. Additionally, the echo profiles are analogous to spectrograms used in audible speech recognition. Therefore, we utilized this network and verified in Sec. 5.4.1 that its recognition accuracy is higher than that of other networks. The detailed description of the network structure can be found in Sec. A.5.

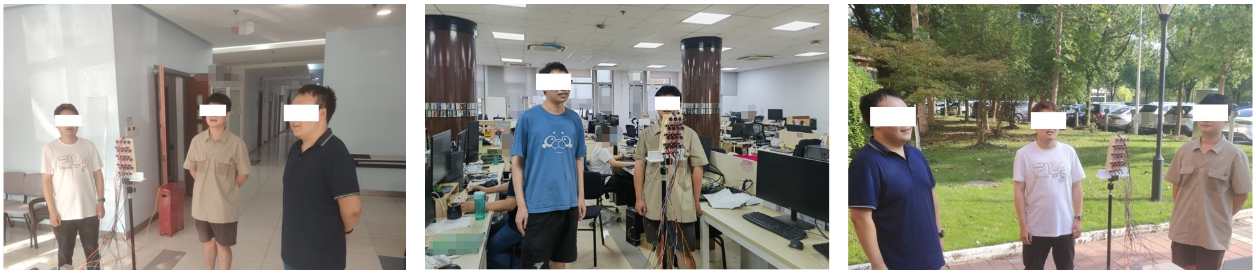
3.4.2 Streaming Recognition. During real-time prediction, M^{2}Silent processes streaming silent speech features using a sliding window. We set the window to $1s$ to capture an entire word, with a $0.15s$ stride to accommodate faster speaking rates (around $100wpm$). Although this approach may cause partial overlap, we address it through data augmentation (Sec. 4.3) by incorporating portions of neighboring words during training.

Due to multiple frames capturing the same word, duplicates are removed based on typical human speech speeds, allowing for legitimate repetition (e.g., strings of numbers). We then enhance recognition accuracy via an N-gram-based correction method [52], which extracts linguistic features (unigram, bigram, trigram, and word posterior probabilities) and feeds them into a CRF [41]. The CRF detects errors and performs corrections by selecting more probable candidates or restructuring sentences.

4 Dataset Construction

4.1 Word Set and Possible Sequences

The word set (Tab. 1) is crucial for accurate intent interpretation. It includes essential action/status words (e.g., "Can," "Need," "Yes," "No"), numerical digits ("Zero" to "Nine"), and common conjunctions/pronouns ("And," "Or," "I," "You"), thereby covering most routine commands and queries.



(a) Uncrowded indoor space.

(b) Crowded indoor space.

(c) Outdoor space.

Figure 7: Different environments of data collection.

Table 1: Word Set

Types	Words
Action/Status	Can, Need, Will, Yes, No, Up, Down, Left, Right, On, Off, Stop, Go, Help, Mute, Unmute, Shut, Volume, Turn, Louder, Lower
Digits	Zero, One, Two, Three, Four, Five, Six, Seven, Eight, Nine
Conjunctions/Pronouns/...	And, Or, But, I, You, It, The, To, Then, Again

Table 2: Possible Sequences

Types	Sequences
Simple Commands	Volume up.
	Turn Left.
	Shut down.
Digital Inputs	Three, one, six, five.
	Eight, six, nine, two, four, one.
Query Sentences	I can help you.
	I need you to shut it off.
	Can you help me turn the volume louder?
	No, try it again.
	Yes, I will unmute it and then turn the volume up.

As shown in Tab. 2, these words can combine into flexible sequences, from simple commands ("Volume up," "Turn left") to more complex requests ("Can you help me turn the volume louder?"). Such sequences may be complete sentences, series of commands, or numerical strings, reflecting the variability of human speech.

4.2 Data Collection.

4.2.1 Participants. The data collection involved 20 participants with a broad demographic distribution. The participant pool consisted of a gender distribution, 9 males and 11 females, and an age range spanning from 18 to 65 years.

4.2.2 Environments. Participants were assigned to 3 different environments to simulate realistic usage scenarios:

- **Uncrowded Indoor Space (Fig. 7(a)):** This environment was a quiet and spacious room, allowing participants to focus on their silent speech interactions without distractions. 6 **participants** were assigned to this environment.
- **Crowded Indoor Space (Fig. 7(b)):** This setting was designed to test the system's performance in a more crowded yet acoustically uncontrolled environment. 8 **participants** were allocated to this environment.
- **Outdoor Space (Fig. 7(c)):** This environment is designed to simulate interactions with M^{2}SILENT used in some outdoor public facilities, such as road alerts, advertisements, and more. The potentially noisy outdoor environment may pose a challenge. 6 **participants** were assigned to this environment.

4.2.3 Data Collection Protocol. Each participant was required to repeat every word from the word set 5 times and produce 200 sequences. These sequences could either be predefined by the experimenters, consisting of combinations of words from the word set, or self-generated by the participants, as long as all words used were from the word set. This approach enabled the collection of both standard and user-generated sequences, enhancing the model's generalization ability. Participants were instructed to speak at a slightly slower-than-average rate to ensure clarity in the silent speech interactions and were encouraged to exaggerate their mouth movements to improve lip-reading accuracy. In multi-user scenarios, up to 4 participants interacted simultaneously with M^{2}SILENT, with each configuration (1 – 4 users) repeated 20 times per setting. Importantly, even in multi-user interactions, the data was processed to treat each word or sequence as a distinct instance for model training, following our feature segmentation method.

4.2.4 Unseen Participants. To further evaluate M^{2}SILENT's performance on unseen users, we recruited an additional 10 participants who were not part of the initial 20. These participants did not contribute to the base model training but were allowed to fine-tune the

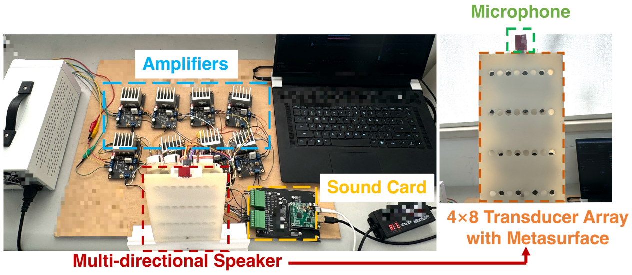

Microphone
Amplifiers
Sound Card
4×8 Transducer Array with Metasurface
Multi-directional Speaker

Figure 8: Hardware prototype setup for M^{2}Silent. This includes a view of the metasurface-embedded parametric array, amplifiers, and the sound card connections.

model using a subset of commands. They were randomly assigned to the three environments and each repeated every word from the word set once and produced 50 sequences.

4.2.5 Data Validation. After data collection, we excluded invalid data points caused by interruptions (e.g., participants stopping midway) or incorrect speech (e.g., mispronouncing a word or sequence). Following this curation process, we obtained a final dataset containing 4207 valid words and 4131 valid sequences.

4.3 Data Augmentation

To enhance the robustness of our silent speech recognition model, we employed several data augmentation techniques.

4.3.1 Warping. Warping involves stretching or compressing the features along the time domain to simulate variations in speech speed, which helps the model learn to recognize the same word or sequence even when spoken more quickly or slowly. We applied warping with coefficients of $[0.5, 2]$, where a coefficient of 0.5 compresses the time axis (simulating faster speech), and a coefficient of 2 stretches it (simulating slower speech).

4.3.2 Shifting. Shifting is applied in two ways: along the time domain and the frequency domain. In **time-domain shifting**, the technique shifts the features along the time axis to simulate different starting times of speech. We used shifting coefficients of $[-1.2, 1.2]$ to create variations where the speech signal starts either earlier or later than usual. For **frequency-domain shifting**, the features are shifted along the frequency axis to simulate variations in the user's distance from the multi-directional speaker. Coefficients of $[-1.5, 1.5]$ were used to represent users being closer or farther from the speaker, respectively.

4.3.3 Noising. Noising involves adding controlled levels of noise to the features and simulating changes in the signal-to-noise ratio (SNR) in the environment. Specifically, we introduced noise with coefficients of $[0.06, 0.08]$.

5 Evaluation

5.1 Evaluation Methodology

5.1.1 Prototype. As depicted in Fig. 8, our prototype is a metasurface-embedded parametric array consisting of 4×8 ultrasonic transducers (Yisheng EU16AOF21H12T [2]) connected to an 8-channel audio source (Lisheng Sound Card [79]). Each channel is powered by a class D amplifier (Texas Instruments OPA541 [30]), which supports up to $50W$ output. The ultrasonic transducers operate at a central frequency of $21kHz$. Each transducer is housed within a metasurface cell, designed as described in [45]. The spacing between the outputs of adjacent channel cells is $8.2mm$, corresponding to half the wavelength of the $21kHz$ signal. Additionally, the single microphone used to receive FMCW echo signals is a MEMS microphone (Analog Devices ADMP404 [17]) with a sampling rate of $96kHz$, positioned centrally at the top of the speaker.

For receiving audio emitted by the multi-directional speaker to measure audio quality, a binaural microphone (Headrec Audio BINAL 2 [7]) with a sampling rate of $96kHz$ is used.

5.1.2 Performance Metrics. We use the following metrics to evaluate M^{2}Silent:

Perceptual Evaluation of Speech Quality (PESQ). PESQ [66], standardized as ITU-T Recommendation P.862 [31], objectively measures speech transmission quality by comparing a reference audio signal to its degraded version. It aligns the signals in time, applies an auditory transform to map them to perceived loudness using psychoacoustic models, and quantifies distortions via symmetric and asymmetric disturbance measures. Audible errors are processed using masking thresholds and aggregated using a nonlinear L_p norm, which emphasizes local distortions. The final score ranging from 1 (poor) to 4.5 (excellent) is computed using the formula:

$$PESQ = 4.5 - 0.1 \cdot d_{\text{sym}} - 0.0309 \cdot d_{\text{asym}}$$

where d_{sym} and d_{asym} are disturbance measures. Scores above 2.5 indicate that the audio can be heard clearly, while scores below 2 indicate very poor audio quality.

Word Error Rate (WER). WER is a standard metric for evaluating speech recognition performance, measuring how closely a system's output matches a reference text. Based on the **Levenshtein distance** [44], it calculates the minimum number of operations, substitutions (S), deletions (D), and insertions (I) to transform one sequence into another. WER is computed as:

$$WER = \frac{S + D + I}{N_w}$$

where N_w is the total number of words in the reference text, which equals $S + D + C$ (C is the number of correctly recognized words). It ranges from 0 (perfect match) to 1 or higher (no similarity or completely incorrect output).

Sequence Error Rate (SER). SER measures the accuracy of predicted sequences in tasks such as speech recognition, where the correct order of words or symbols is critical. Unlike WER, which focuses on individual words, SER assesses errors in the entire sequence, and it is computed as:

$$SER = \frac{N_{is}}{N_s}$$

where N_{is} is the number of incorrect sequences and N_s is the total number of sequences.

5.2 Overall Performance

We use WER, SER, and PESQ to evaluate the overall performance of M^{2}Silent with different numbers of users (from 1 to 4). The

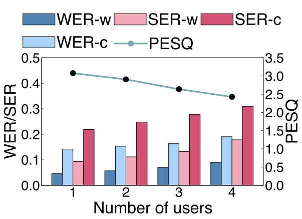


Figure 9: Overall Performance: WER, SER, PESQ under different numbers of users.

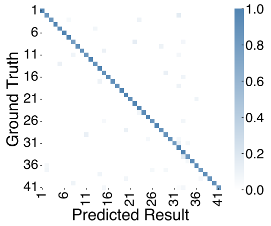


Figure 10: Confusion matrix of word-level recognition across the 41 words in the word set.

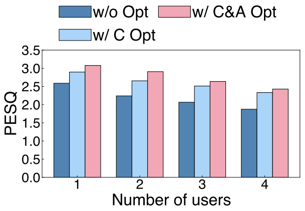


Figure 11: Ablation Study 1: impact of optimization strategies on audio quality.

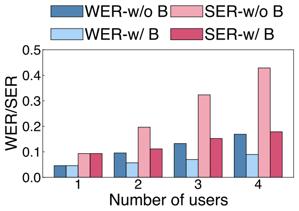


Figure 12: Ablation Study 2: impact of blind source separation on recognition.

results are shown in Fig. 9. PESQ decreases as the number of users increases. Initially, for a single user, the PESQ score is 3.02. However, as the number of users increases, the PESQ score decreases to 2.43 when there are 4 users. This indicates that the perceived speech quality decreases because the system has to handle more users simultaneously, but it also shows that using FMCW signals as carriers is fully capable of supporting multiple users.

For silent speech recognition performance, we use WER and SER under within-user (w) and cross-user (c) conditions. WER-w and WER-c represent the error rates within and across users, respectively, while SER-w and SER-c reflect sequence error rates for these conditions. Cross-user performance is evaluated using a fine-tuned model based on the 10 words from the users in 4.2.4. WER-w remains low, ranging from 4.56% to 8.96% as the number of users increases, while WER-c is higher, between 14.22% and 19.05%. This indicates better performance with within-user data than across users, suggesting that more words for fine-tuning could improve accuracy, though it may also complicate recognition for new users. For sequence testing, SER-w and SER-c show trends similar to WER, with SER-w ranging from 9.32% to 17.85%, and SER-c increasing from 21.87% to 30.96% as the number of users grows. This highlights that sequence errors become more prominent with multiple users. Overall, SER is acceptable, with an average error occurring once in five complete conversations, which meets the needs of most users in silent speech interactions.

In addition, we explored the accuracy of each word from all within-user and cross-user test sets. From the confusion matrix shown in Fig. 10, the mean accuracy of each word is 92.13%, and the standard deviation is 6.49%. This shows that the model can recognize each word accurately. However, for some words with very similar pronunciation patterns, such as "on" and "or", which have short durations and similar mouth shapes, the model may make mistakes, but such mistakes can have a chance to be corrected by grammar-based error correction mentioned in Sec. 3.4.2.

5.3 Ablation Study

5.3.1 Impact of Optimization Strategies. In the first ablation study, we assess the impact of optimization strategies on audio quality using FMCW signals as carriers, and the study compares three configurations as shown in Fig. 11. "w/o Opt" refers to the baseline scenario where no optimization is applied. Here, we use a linear sawtooth waveform as the FMCW signal, with a chirp length of 0.1 seconds and a bandwidth of 4 kHz. The results show that this configuration provides relatively low PESQ scores in all user cases, with the worst case being a PESQ of 1.87 for 4 users, which is insufficient to support multiple users. "w/ C Opt" is a configuration that optimizes the carrier signal through an enumeration search. As mentioned in Sec. A.2, we tried various types of FMCW waveforms, different bandwidths, and chirp lengths. We eventually selected a linear triangular waveform with a chirp length of 0.25 seconds and a bandwidth of 2 kHz. This configuration ensures both audibility and good perceptual performance (as a narrower bandwidth would increase the ambiguity of cross-correlation). Compared to the baseline, this optimization improves PESQ by about 0.4 in all user cases. "w/ C&A Opt" is a setting that further addresses non-linear distortion and reduces audio interference caused by time-varying FMCW signals. PESQ in this configuration improves further by 0.2, indicating that optimizing both the carrier and audio can significantly improve audio quality. The streaming processing of the optimization has almost no impact on system latency, as the optimizations take only 0.04 seconds to process a 5-second audio file.

Regarding user experience and system impact, the FMCW signal parameters we provide allow users to achieve an auditory experience almost identical to that of a standard speaker while maintaining high recognition accuracy. If the FMCW signal uses a narrower bandwidth or a longer chirp duration, the FMCW becomes smoother, further enhancing the user's auditory experience. However, this also increases the ambiguity in the cross-correlation process, decreasing recognition accuracy. Therefore, using the FMCW signal optimization results we provide to balance auditory experience and recognition accuracy is recommended.

5.3.2 Impact of Blind Source Separation. In the ablation study 2, we investigate the impact of blind source separation (BSS) on recognition accuracy and evaluate the WER and SER for different numbers of users. All evaluated here are within-user, as shown in Fig. 12. We compare 2 configurations: w/o B: without BSS, w/ B: with BSS. When there are multiple users, the WER is generally lower when BSS is enabled. For example, for 2 users, WER-w/ B is about 5.69%, while WER-w/o B is about 5.46%. This trend continues as the number of users increases. In the case of 4 users, WER-w/o B increases significantly to about 16.88%, while WER-w/ B remains at 8.94%. Similarly, SER is lower when BSS is applied. For example, SER-w/ B is lower than SER-w/o B at all numbers of users. In the case of 4 users, SER-w/o B rises sharply to 42.83%, indicating a significant

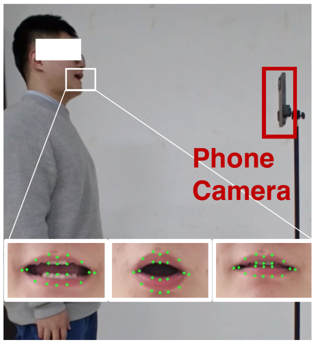


Figure 13: Visual input: using the camera on the phone for silent speech recognition.

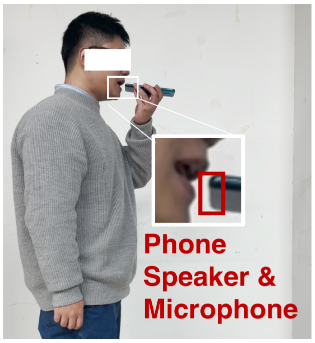


Figure 14: Phone Speaker & Microphone: using the speaker and microphone of the phone for silent speech recognition.

drop in sequence recognition without BSS, where the model is essentially unable to perform silent speech recognition, while SER-w/B remains below 20%.

BSS can significantly improve the accuracy of silent speech recognition in environments with many users. However, in a single-user scenario, it has no impact on accuracy. Our system is designed to serve multiple users, so BSS enables the system to accurately recognize speech even when many users are interacting simultaneously. This prevents users from having to repeat their silent speech multiple times, thus enhancing convenience. Additionally, the BSS algorithm operates quickly, at the millisecond level, and performs well on systems with average computing power.

5.4 Comparative Study

5.4.1 Comparison of Different Machine Learning Models. In this comparative study, we analyzed the performance of various models in terms of WER and SER. Tab. 3 shows the results of five different models: LSTM [76], GRU [68], DS-CNN [103], Attention ResNet [38], and SilentMatch. It can be found that SilentMatch has the best performance, with the lowest WER of 6.52% and the lowest SER of 12.81%. SilentMatch outperformed all other models, indicating that its architecture is very effective in minimizing word and sentence errors. The reason is that SilentMatch uses a scalable one-dimensional time channel separable convolutional neural network designed for word recognition. It is robust to background noise and has a small number of parameters, making it compact in devices with limited computing resources.

5.4.2 Comparison of Different Silent Speech Recognition Schemes. We compared the silent speech recognition capability of M^{2}Silent with methods leveraging visual input [75] and traditional mobile speaker-microphone setups [98]. For the visual input (Fig. 13), we used the rear camera of an iPhone 15 Pro Max to record the speaker at a distance of $1-2$ meters. Using facial landmark detection algorithm [36] provided by Dlib [40], we identified the facial key points of the speaker, cropped the mouth region, and employed an end-to-end network [75] for recognition. In the phone speaker and microphone setup (Fig. 14), we utilized the bottom speaker of a Redmi 10X to emit a multi-frequency continuous wave signal ranging from $18kHz$ to $22kHz$. The speaker brought their mouth close to the bottom of the phone to produce silent speech. The microphone captured the signal, extracting its phase and amplitude, which was then processed using a hierarchical convolutional neural network [98] for recognition. When identifying sequences, we employed the sliding window method.

We collected data from 5 participants under both schemes, with an additional 3 participants as unseen users. The data collection protocol followed that of M^{2}Silent. As shown in Tab. 4, we compared WER and SER, averaged across within-user and cross-user scenarios. The results indicate that M^{2}Silent delivers comparable performance to the mobile speaker-microphone-based method, demonstrating reasonable accuracy in silent speech recognition. However, its performance lags behind the visual input-based method, which directly captures lip movements for a more intuitive representation. Nevertheless, considering that users are more tolerant of errors in silent speech recognition tasks [59], and accounting for the visual method's sensitivity to environmental lighting and privacy concerns, the recognition capability of M^{2}Silent is deemed acceptable.

5.5 Sensitivity Study

In the sensitivity study, the 20 initial participants and 10 unseen participants from Sec. 4.2 were involved, and we additionally invited 10 more unseen participants, consisting of 7 males and 3 females, aged $23-54$ years, with an average age of 32.3 years. The study reveals how various factors impact the accuracy of the system's silent speech recognition and audio quality.

5.5.1 Impact of Angles. We evaluated the impact on audio and silent speech recognition when users stand at different angles. Using a protractor, we measured the angle range around M^{2}Silent and asked users to stand approximately 1.5 meters away at different angles. As shown in Fig. 15(a), as the angle increases from 0° to 80°, WER increases slightly, while SER rises more significantly, especially at larger angles. This is due to the directional speaker's volume attenuation at wider angles during beamforming, which weakens the sensing signal and reduces the signal-to-noise ratio. PESQ drops from 2.95 to 2.53, indicating this attenuation in perceived audio. However, PESQ scores above 2.5 still allow users to hear clearly, and extreme user positioning is rare, so users can be reminded to adjust their position if needed.

Users experience effective bidirectional interaction within $\pm 60^\circ$ from M^{2}Silent. While full-directional interaction is not supported, it is adequate for most user interactions, as users generally engage within a $\pm 30^\circ$ range in front of the speaker [23].

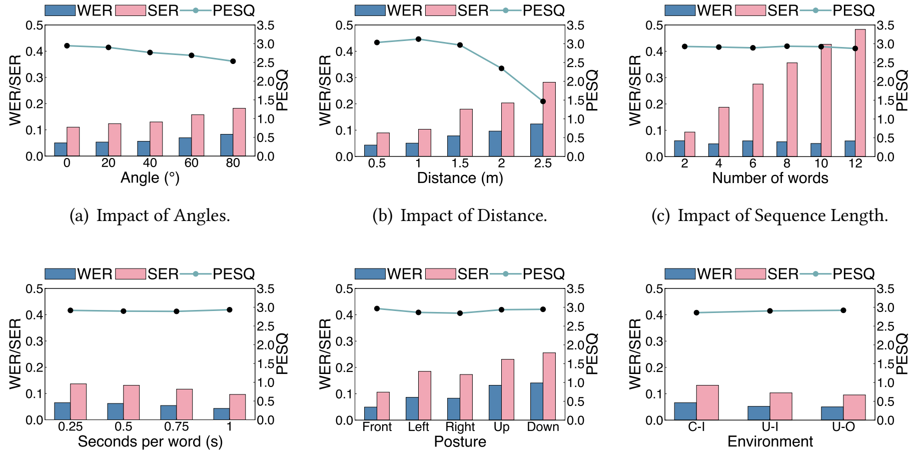
5.5.2 Impact of Distance. We evaluated the effect of user distance from M^{2}Silent. Using a tape measure, we marked different distances along a straight line opposite M^{2}Silent and asked users to stand at these marks while performing silent speech. As shown in Fig. 15(b), as the distance increases from $0.5m$ to $2.5m$, both WER and SER increase, indicating a decline in speech recognition accuracy with distance. This effect is noticeable beyond $2m$, as ultrasonic waves attenuate rapidly, weakening the reflected sensing signal. PESQ also decreases with increasing distance, from around 3.0 at $0.5m$ to about 2.0 at $2.5m$, reflecting a decline in audio quality.

Table 3: Performance comparison with different models.

Model	WER	SER
LSTM	10.55%	23.33%
GRU	9.69%	21.65%
DS-CNN	9.32%	21.41%
Attention ResNet	7.33%	15.16%
SilentMatch	**6.52%**	**12.81%**

Table 4: Performance comparison with different silent speech recognition schemes.

Scheme	WER	SER	Long Distance	Privacy	Dark
Visual Input	4.12%	8.36%	✔	✘	✘
Phone Speaker & Microphone	8.26%	17.58%	✘	✔	✔
M^{2}Silent	6.92%	13.34%	✔	✔	✔



(a) Impact of Angles.

(b) Impact of Distance.

(c) Impact of Sequence Length.

(d) Impact of Speaking Speed.

(e) Impact of Postures.

(f) Impact of Environments.

Figure 15: Sensitivity Analysis.

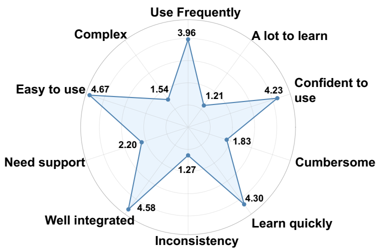


Figure 16: System Usability Scale (SUS) results from user evaluation. This figure shows the participants' responses to different metrics, including ease of use, integration, confidence, and complexity of M^{2}Silent.

Users within $2m$ of the device can enjoy a good interaction experience, with PESQ around 2.5 and WER around 9.5%. This result aligns with the optimal viewing distance in many scenarios, such as viewing artworks in a museum ($1.49m - 2.12m$ [12]) and the distance between an interactive interface and a car seat (generally within $0.7m$ [57]).

5.5.3 *Impact of Sequence Length.*

Longer sequences lead to an increase in SER. Fig. 15(c) shows that SER rises sharply when the sequence reaches 12 words. A possible method to reduce SER is introducing a transformer model, which would significantly increase the required training data. However, WER remains relatively stable across different sequence lengths. PESQ remains unchanged, but recognition performance declines as sequences become longer.

Most voice commands are short, typically within 8 words [19], which suffices for users to say a few keywords for a command. In these cases, M^{2}Silent can complete the recognition with an average performance of 5.62% WER and 22.81% SER, which is in line with user expectations, because most speech recognition systems also process short commands individually [48, 88].

5.5.4 *Impact of Speaking Speed.*

As shown in Fig. 15(d), changes in speaking speed (measured in seconds per word) have little effect on WER and SER, with both remaining usable at different speeds. However, slowing down the speaking speed slightly reduces WER and SER, as this produces more pronounced facial movements.

The general speaking speed is approximately $0.3 - 1$ seconds per word [25]. At this speaking speed, M^{2}Silent can maintain a silent speech recognition performance of 6.36% WER and 13.02% SER, so users can comfortably use M^{2}Silent at a normal speaking pace.

5.5.5 *Impact of Postures.*

Different postures, including facing forward and tilting the head left, right, up, or down, result in significant variations in SER, as shown in Fig. 15(e). A downward posture leads to the highest error rate. WER also increases slightly with changes in posture. PESQ remains stable across all postures, as posture does not typically influence perceived audio. However, differences in ear volume due to posture might be introduced.

In several cases, users may not be directly facing M^{2}Silent. For example, in a car, shaking caused by driving may occur, or in a museum, users may observe artwork without directly facing M^{2}Silent, potentially resulting in slight facial deviations. However, these deviations are generally tolerable, with a WER of 8.48% and an SER of 17.92% in such cases. While in scenarios involving looking up or down, the WER reaches 13.67% and the SER 24.32%, users are unlikely to excessively tilt their heads up or down in most situations, as doing so would make it inconvenient to listen to M^{2}Silent's audio output, prompting them to adjust their posture naturally.

Table 5: User feedback and responses.

Feedback	Response
1. *"When silent speech is not properly recognized, the system should provide clearer feedback. Offering more immediate system feedback during use can help users understand mistakes and adjust their speech patterns accordingly, enhancing the user experience."*	Yes, if the user speaks too quickly, the system can issue targeted reminders to slow down.
2. *"Does prolonged exposure to ultrasonic waves negatively affect human hearing?"*	Our ultrasonic wave intensity meets international standards, and we can implement activation steps to reduce potential long-term impacts.
3. *"If I'm not directly facing the speaker, does the recognition accuracy decrease, and is there a way to avoid this?"*	We recommend users face the speaker when performing silent speech. We've also tested training the model with data collected from various postures, though this may introduce additional complexity.
4. *"Can I move while speaking?"*	The tracking feature can be achieved by sensing motion through FMCW signals, but this may disrupt the system's continuity. The system can be adapted to support movement, but users will have a better experience when stationary.

5.5.6 Impact of Environments. As shown in Fig. 15(f), we tested three different environments: crowded indoor (C-I), uncrowded indoor (U-I), and uncrowded outdoor (U-O). WER and SER were the lowest in the uncrowded outdoor environment, while both error rates increased in more uncontrolled environments. PESQ remained relatively stable across different environments, with better audio quality in uncrowded outdoor settings. Overall, multipath effects indoors caused slight interference.

Although in indoor environments and crowded spaces, the presence of multipath effects may slightly reduce the signal-to-noise ratio of received sensor signals, the impact is generally minimal. Compared to open spaces, under crowded conditions, the PESQ decreases by 0.04, the WER increases by 1.56%, and the SER increases by 3.32%. Since these environmental factors are usually static, the differential method mentioned in Sec. A.3 can mitigate their effects, ensuring that the user experience remains unaffected. This demonstrates that M^{2}Silent is fully capable of operating effectively in potentially crowded spaces, such as cars with seats or exhibition halls with numerous displays.

5.6 User Study

5.6.1 System Usability Scale. This study used the System Usability Scale (SUS) to evaluate user interactions with M^{2}Silent in different environments. SUS is a reliable tool that measures usability through a standardized set of 10 questions, each rated on a 5-point Likert scale. These questions assess the system's ease of use, complexity, and user confidence.

The main findings from the SUS analysis (Fig. 16) indicate that participants found the system suitable for frequent use (scoring 3.96), with minimal impact from the ultrasonic waves. They considered the system easy to use (scoring 4.67) because there was no need for manual adjustments to the multi-directional speakers, as the beams automatically aligned with them. Participants felt the system was well-integrated (scoring 4.58), as the coordination between the speakers and microphones made two-way communication convenient. The system's ease of learning received a score of 4.3, as directly speaking lip movements was more convenient than learning additional gestures. Participants also expressed confidence in using the system (scoring 4.23), as it made open-voice interaction more comfortable in privacy-sensitive or quiet environments. For some issues, such as complexity, need for support, and inconsistency, user feedback and responses were summarized in Tab. 5. These insights highlight both the strengths of the system in terms of usability and areas for further improvement to enhance the user experience.

5.6.2 Social Acceptance. We explored the social acceptance of M^{2}Silent from the perspectives of key stakeholders, interviewing two car designers, one museum manager, one banker, and five general users. Among them, two car designers and three general users experienced M^{2}Silent in person, while the others watched a remote online demonstration. Their ages ranged from 21 to 48. We gathered their comments on the system's acceptability in Tab. 6.

Two car designers expressed that M^{2}Silent could be implemented in vehicles. However, one car designer raised concerns about the added cost of incorporating an additional audio system in vehicles. The museum manager considered M^{2}Silent an ideal solution for enabling visitors to interact with exhibits without disturbing others. The system's ability to silently inquire about directions or exhibit details aligns with the goal of maintaining a contemplative atmosphere. The banker viewed M^{2}Silent as a breakthrough for safeguarding confidentiality, particularly when discussing contracts and transactions. Clients could silently convey sensitive information, but the banker expressed concerns about its potential to replace other methods. Younger users appreciated the ability to silently interact with devices in open spaces, finding it useful for activities like controlling music at parties or managing devices without disturbing others. In contrast, older users emphasized the need for guidance to become familiar with the system.

In summary, M^{2}Silent addresses concerns about equipment cost and ease of use. The device incurs minimal costs and can replace some existing speaker and microphone systems. Additionally, offering more detailed instructions, such as having directional speakers explain the silent speech function, could improve user-friendliness.

Table 6: Comments from several participants in social acceptance interviews.

Participant	Comments
Car Designer 1	*"The automotive industry might accept the system because it enables quiet and personalized interactions within vehicles, aligning with modern design needs for undisturbed and serene in-car environments."*
Car Designer 2	*"It might require reducing some existing speakers."*
Museum Manager	*"This system provides a futuristic and respectful solution for visitor engagement."*
Banker	*"While this method is intuitive, whether it can fully replace some button-based services requires further evaluation. It could, however, be more user-friendly for individuals with limited hand mobility."*
User 1	*"It's like having a secret voice assistant that no one else can hear."*
User 2	*"If someone shows me how to use it, that would be wonderful."*

6 Discussion

6.1 Use Cases

Voice-based services in open scenarios, such as in-car navigation or museum audio guides, are irreplaceable. Research has shown that voice interaction is more convenient and intuitive than tapping on a screen [59]. In these multi-user scenarios, each user has unique content needs and interaction requests. While personalized audio services can also be delivered through personal devices like headphones, phones, or watches, these options are often rejected due to discomfort, aesthetic concerns, occupied hands, or high costs. Our system enables each user to hear different content without interference and contact-free interaction in multi-user settings. Additionally, enhancing multi-directional speakers with silent speech recognition capabilities is valuable, allowing for bidirectional communication without additional devices like cameras, as silent speech recognition has already been proven to be an efficient interaction method that most users can accept.

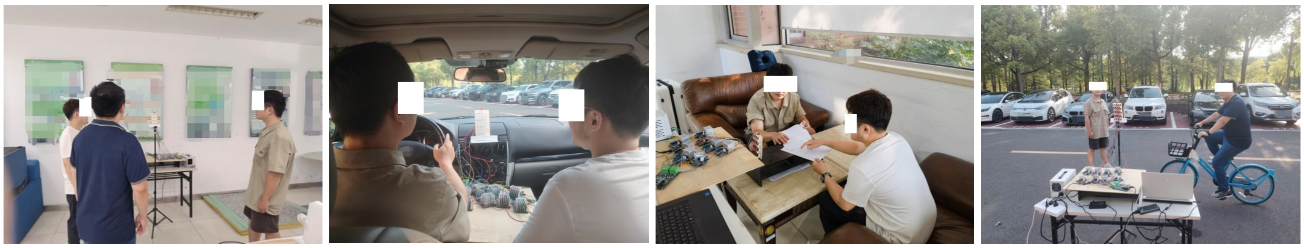
The added value of our system lies in integrating multi-user silent speech recognition into multi-directional speakers, which can play a crucial role in shared environments where interactions may require 1) maintaining a **quiet** environment, 2) dealing with **noisy** surroundings, 3) avoiding user **embarrassment**, or 4) addressing **privacy** concerns. As shown in Fig. 17, M^{2}Silent can be applied in a variety of real-world scenarios:

In exhibition rooms (Fig. 17(a)): M^{2}Silent enables visitors to silently inquire about detailed information regarding exhibits and receive responses through directional sound waves, without disturbing others or compromising the **quiet** environment. Many museums worldwide have adopted noise standards [5, 18, 28, 80], prohibiting the use of loudspeakers and requiring visitors to refrain from speaking loudly [8, 29, 56]. However, in such settings, voice interaction can provide significant convenience, such as asking for details about exhibits or the location of items. Unfortunately, using traditional loudspeakers and human voice communication violates the silence policies, and very few visitors are willing to purchase or rent additional devices like audio guides [27, 35]. M^{2}Silent serves as an ideal solution to enhance visitor experiences in quiet, shared environments. Furthermore, many users are hesitant to use voice interaction in public due to **embarrassment** [9, 21, 50, 81], fearing that others might judge their requests as trivial or silly, like asking questions with obvious answers. In such scenarios, M^{2}Silent allows users to interact silently, avoiding embarrassment and making voice interaction more acceptable to them.

In-vehicle scenario (Fig. 17(b)): M^{2}Silent provides personalized audio interactions for car occupants while maintaining a **non-interfering** and mutually comfortable environment, allowing the driver to focus on navigation sounds without being disturbed by music played by other passengers. This is crucial, as additional noise makes it harder for the driver to hear important sounds, such as alarms, horns, or the vehicle's own alerts, thus increasing the likelihood of accidents [33, 70, 83]. Additionally, the acoustic interface provided by M^{2}Silent is superior to touch-based interactions, preventing driver distraction. The quiet, non-interfering interaction that M^{2}Silent enables for each passenger allows some to rest comfortably, as in-car noise can create an uncomfortable environment, leading to stress or fatigue. Such noise has a negative impact on health, causing stress and sleep disorders [6, 51]. Moreover, the prevalence of ride-hailing services has increased **privacy** concerns in cars [1, 61, 104]. Drivers may not want to discuss the personal information in front of unfamiliar passengers. M^{2}Silent ensures the privacy of in-car interactions by enabling the driver to use silent speech and directional speakers, preventing other passengers from overhearing. This allows the driver to use hands-free calling with greater confidence.

In transactions (Fig. 17(c)): M^{2}Silent can protect sensitive information and help ensure **privacy** in public spaces like banks. Numerous reports highlight that speaking sensitive information aloud (e.g., personal details) may lead to eavesdropping or exposure of private addresses [15, 55, 78, 90]. It has been pointed out that discussing personal information of service providers in public settings can easily violate confidentiality agreements [72]. By using M^{2}Silent, users can discreetly convey key details without being overheard, providing a high level of privacy in environments such as banks or private offices. For instance, in a bank, if the user is writing by hand or has a hand disability that makes button use difficult, M^{2}Silent can replace button inputs to enter sensitive information. Another example is in the office, where a trader needing to complete a transaction can use silent voice input of M^{2}Silent to enter the transaction amount, account details, and password. The staff can then replay the customer's information and provide private feedback, such as confirming the accuracy of sensitive details.

On the street (Fig. 17(d)): M^{2}Silent can direct important information to pedestrians on the street and allow for interaction, even in **noisy** outdoor environments. In such settings, users may have to shout loudly for a voice system to recognize their requests, such



(a) In the exhibition room. (b) In the car. (c) In the transaction. (d) On the street.

Figure 17: Real world use cases.

as quickly changing a traffic light or asking for contact information from an advertiser [60, 82, 96]. A company has placed self-service machines on the street to allow people to customize items through voice commands [49], but such systems often struggle with recognition in noisy environments. Our system enables users to interact efficiently even in high-noise environments. For example, M^{2}Silent can alert users needing traffic updates, such as warnings about oncoming vehicles, while other pedestrians can use silent voice commands to inquire about advertising details unaffected by surrounding traffic noise. In outdoor environments, using silent voice input also avoids the **embarrassment** of speaking loudly. A study has shown that although 90% of people have tried voice interaction, only 6% have used it in outdoor public spaces [3]. The introduction of M^{2}Silent will make street services more efficient and, compared to traditional voice services, allow more users to engage with them.

6.2 Time Delay and Resource Cost

The primary time delay in M^{2}Silent arises from audio output optimization and silent speech recognition. Audio optimization takes 0.032 seconds per second of audio, while silent speech recognition requires 0.074 seconds per second of input. This minimal delay ensures near-instantaneous, real-time responses, making the system highly efficient for everyday use. Additionally, M^{2}Silent is lightweight ($8cm$ x $18cm$) and affordable, priced at $352 USD.

6.3 User Tracking

In scenarios where the user is moving or in a car that vibrates a lot, M^{2}Silent may need to track the user's head position to interpret lip movements. This can be achieved through beam scanning, and because M^{2}Silent inherently emits FMCW signals, it can sense the user's position. Potentially, the system can ask users whether they wish to initiate interaction in order to activate the device. By extending the FMCW signal to incorporate real-time user tracking, the system can dynamically adjust based on the user's position and movements.

6.4 Health Concerns

Our research has received approval from the IRB. In this work, we implemented M^{2}Silent using transducers operating at a central frequency of $21kHz$. The transmission power complies with FDA safety standards, which stipulate that the sound level for $21kHz$ should not exceed $80dBSPL$ (decibels Sound Pressure Level) at a distance of $1m$. Note that directional speakers can use ultrasonic signals of different frequencies as carriers, and the safety standards vary depending on the frequency of the ultrasonic signal. For instance, a frequency of $40kHz$ can reach $120dBSPL$, while lower frequencies result in lower sound pressure levels. Although the sound pressure level of $21kHz$ is $80dBSPL$ at $1m$, the audio demodulated from it can be clearly heard by the human ear ($60dBSPL - 65dBSPL$). In future work, it is worth considering replacing the current setup with higher-frequency ultrasonic transducers to achieve greater sound wave emission energy and support longer distances.

The current prototype is designed for use at moderate distances (this fully aligns with scenarios such as inside a car, in museums, etc.), ensuring that ultrasound exposure remains within safe limits. By adhering to FDA guidelines, we can ensure the system's safety during prolonged use, even in proximity. If users have additional concerns about the system's safety, we believe that adding an activation feature to M^{2}Silent, which would only generate ultrasound during interactions, is feasible.

7 Limitations & Future Work

In this paper, we propose a new prototype for achieving acoustic-based multi-user, bidirectional silent interaction in open scenarios. However, there are still limitations when dealing with complex real-world environments. We discuss these limitations and envision future work to address them.

Longer distance. The maximum distance supported by M^{2}Silent for silent voice interaction is around 2 meters. While this range is suitable for most scenarios, such as inside a car or in museums (where the optimal viewing distance for artwork is $1.49m$ – $2.12m$ [12]), its performance may decrease for users standing at a greater distance. Future work could explore increasing the number of ultrasonic transducers to enhance transmission power and strategically deploying distributed M^{2}Silent units in the environment to achieve broader coverage.

More users. M^{2}Silent performs well when interacting with up to three users simultaneously. However, performance degrades when supporting four or more users. This is because additional users require M^{2}Silent to emit more beams. Since the beam has a certain width, this causes beam overlap and may cause confusion. Future work could focus on optimizing the spacing and arrangement of ultrasonic transducer arrays to produce more precise beams and avoid overlap, enabling M^{2}Silent to support a greater number of users simultaneously.

Occlusion. In crowded environments, such as indoor spaces with many pillars, occlusion may occur. These can impact both the user's ability to hear sound and the accuracy of silent voice recognition. To address this problem, M^{2}Silent could leverage potential reflectors in the environment to bypass obstacles and communicate with users. Alternatively, the problem can be avoided by flexibly deploying multiple M^{2}Silent units within the space.

8 Conclusion

In conclusion, M^{2}Silent introduces a novel approach for enabling multi-user silent speech interaction in shared spaces. By combining multi-directional speakers, FMCW signal processing, and deep learning-based speech recognition, the system achieves high accuracy in recognizing silent speech while maintaining privacy and minimizing sound leakage. The system's ability to simultaneously support multiple users in environments such as cars, museums, and outdoor settings highlights its versatility and practicality. The low latency and minimal resource cost further ensure a seamless user experience, making M^{2}Silent a valuable solution for real-world applications where privacy and silent interaction are essential.

Acknowledgments

We would like to express our gratitude to all anonymous reviewers for their constructive comments. This work is supported in part by National Natural Science Foundation of China (No. 61936015), Natural Science Foundation of Shanghai (No. 24ZR1430600) and Shanghai Key Laboratory of Trusted Data Circulation and Governance, and Web3.

References

[1] Ransford A Acheampong. 2021. Societal impacts of smart, digital platform mobility services—an empirical study and policy implications of passenger safety and security in ride-hailing. *Case Studies on Transport Policy* 9, 1 (2021), 302–314.

[2] AliExpress. 2024. EU16AOF21H12T. https://www.aliexpress.us/item/3256801485229376.html

[3] AlphaTech. 2023. Voice search embarrassment, at least 20% refuse to use voice search in public. https://alphatech.technology/Blog-Entry-srk/Voice-search-embarrassment-bek

[4] Carl Andersson and Jens Ahrens. 2018. A method for simultaneous creation of an acoustic trap and a quiet zone. In *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*. IEEE, IEEE, Hoes Ln Piscataway, NJ, 622–626.

[5] ASI Architectural. 2020. Acoustics in Museums: The Science of Sound. https://www.asiarchitectural.com/acoustics-in-museums.

[6] Ane Arregi, Oscar Vegas, Aitana Lertxundi, Ana Silva, Isabel Ferreira, Ainhoa Bereziartua, Maria Teresa Cruz, and Nerea Lertxundi. 2024. Road traffic noise exposure and its impact on health: evidence from animal and human studies—chronic stress, inflammation, and oxidative stress as key components of the complex downstream pathway underlying noise-induced non-auditory health effects. *Environmental Science and Pollution Research* 31, 34 (2024), 46820–46839.

[7] Headrec Audio. 2024. BINAL 2 Binaural Microphone. https://headrec.com/products/binal-two

[8] Rabi Chislon BANTAI. 2021. REGULATION OF LIBRARY NOISE POLICY FOR EFFECTIVE NOISE CONTROL. *Regulation* 3, 1 (2021), 1–10.

[9] Evan Bartlett. 2024. Voice Assistants are terrible. https://www.evbart.com/voice-assistants-are-terrible

[10] Braeden C Benedict, Mohammad Meraj Ghanbari, and Rikky Muller. 2022. Phased array beamforming methods for powering biomedical ultrasonic implants. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 69, 10 (2022), 2756–2765.

[11] Anne-Claire Bourland, Peter Gorman, Jess McIntosh, and Asier Marzo. 2017. Project telepathy: Targeted verbal communication using 3D beamforming speakers and facial electromyography. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1508–1515.

[12] Claus-Christian Carbon. 2017. Art perception in the museum: How we spend time and space in art exhibitions. *i-Perception* 8, 1 (2017), 2041669517694184.

[13] Heng-Jui Chang, Alexander H Liu, Hung-yi Lee, and Lin-shan Lee. 2021. End-to-end whispered speech recognition with frequency-weighted approaches and pseudo whisper pre-training. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, IEEE, Hoes Ln Piscataway, NJ, 186–193.

[14] Seungjin Choi, Andrzej Cichocki, Hyung-Min Park, and Soo-Young Lee. 2005. Blind source separation and independent component analysis: A review. *Neural Information Processing-Letters and Reviews* 6, 1 (2005), 1–57.

[15] Corporate Communications. 2024. You're a Corporate Communications professional. How do you keep sensitive information confidential? https://www.linkedin.com/advice/1/youre-corporate-communications-professional-zokmf

[16] David G Crighton. 1979. Model equations of nonlinear acoustics. *Annual Review of Fluid Mechanics* 11, 1 (1979), 11–33.

[17] Analog Devices. 2012. ADMP404 MEMS Microphone. https://www.analog.com/media/en/technical-documentation/obsolete-data-sheets/ADMP404.pdf

[18] Dario D'Orazio, Federico Montoschi, and Massimo Garai. 2020. Acoustic comfort in highly attended museums: A dynamical model. *Building and Environment* 183 (2020), 107176. https://doi.org/10.1016/j.buildenv.2020.107176

[19] Jean-Yves Fourniols, Nadim Nasreddine, Christophe Escriba, Pascal Acco, Julien Roux, and Georges Soto-Romero. 2018. An overview of basics speech recognition and autonomous approach for smart home IOT low power devices. *Journal of Signal and Information Processing* 9, 4 (2018), 239.

[20] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. Echowhisper: Exploring an acoustic-based silent speech interface for smartphone users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–27.

[21] Nielsen Norman Group. 2020. Intelligent Assistants: Creepy, Childish, or a Tool? Users' Attitudes Toward Alexa, Google Assistant, and Siri. https://www.nngroup.com/articles/voice-assistant-attitudes

[22] Đorđe T Grozdić and Slobodan T Jovičić. 2017. Whispered speech recognition using deep denoising autoencoder and inverse filtering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 12 (2017), 2313–2322.

[23] Robert Harley. 2019. The Four Secrets of Speaker Placement. https://www.theabsolutesound.com/articles/the-four-secrets-of-speaker-placement

[24] Robin Hofe, Stephen R Ell, Michael J Fagan, James M Gilbert, Phil D Green, Roger K Moore, and Sergey I Rybchenko. 2013. Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. *Speech Communication* 55, 1 (2013), 22–32.

[25] Lan-Fen Huang and Tomáš Gráf. 2020. Speech Rate and Pausing in English: Comparing Learners at Different Levels of Proficiency with Native Speakers. *Taiwan Journal of TESOL* 17, 1 (2020), 57–86.

[26] Victor F Humphrey, Stephen P Robinson, John D Smith, Michael J Martin, Graham A Beamiss, Gary Hayman, and Nicholas L Carroll. 2008. Acoustic characterization of panel materials under simulated ocean conditions using a parametric array source. *The journal of the acoustical society of America* 124, 2 (2008), 803–814.

[27] Guide ID. 2024. The Audio Guide Pricing Dilemma. https://www.guide-id.com/explore/the-audio-guide-pricing-dilemma

[28] INFOGRAPHIC. 2024. Acoustics in the Modern Museum. https://www.baswana.com/news/museum-acoustics.

[29] The Franklin Institute. 2023. Museum Policies. https://fi.edu/en/museum-policies

[30] Texas Instruments. 2016. OPA541. https://www.ti.com/lit/ds/symlink/opa541.pdf

[31] International Telecommunication Union. 2021. P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. https://www.itu.int Retrieved 2021-04-20.

[32] Jeeyeon Jeong, Yaeri Kim, and Taewoo Roh. 2021. Do consumers care about aesthetics and compatibility? The intention to use wearable devices in health care. *SAGE Open* 11, 3 (2021), 21582440211040070.

[33] Tao Jin, Xiaoxu Liu, Chunpeng Chen, Yuting Xia, Xinyu Liu, Meiyu Lv, and Li Li. 2024. The impact of environmental noise on drivers' cognitive abilities: A case study on in-vehicle voice interaction interfaces. *Applied Ergonomics* 117 (2024), 104247.

[34] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: " Hearing" Your Silent Speech Commands In Ear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–28.

[35] My Smart Journey. 2021. 6 Reasons Why Audio Guides Are Changing in 2021. https://mysmartjourney.com/en-ca/post/6-reasons-why-audio-guides-are-changing-in-2021

[36] Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Hoes Ln Piscataway, NJ, 1867–1874.

[37] Soundlazer kickstarter. 2016. Soundlazer. https://www.kickstarter.com/projects/richardhaberkern/soundlazer.

[38] Byeonggeun Kim, Simyung Chang, Jinkyu Lee, and Dooyong Sung. 2023. Broadcasted Residual Learning for Efficient Keyword Spotting. arXiv:2106.04140 [cs.SD] https://arxiv.org/abs/2106.04140

[39] Byung-Chul Kim and I-Tai Lu. 2000. Parameter study of OFDM underwater communications system. In *OCEANS 2000 MTS/IEEE Conference and Exhibition. Conference Proceedings (Cat. No. 00CH37158)*, Vol. 2. IEEE, IEEE, Hoes Ln Piscataway, NJ, 1251–1255.

[40] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research* 10 (2009), 1755–1758.

[41] John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*. Williamstown, MA, ACM, New York, NY, USA, 3.

[42] Dominic Langlois, Sylvain Chartier, and Dominique Gosselin. 2010. An introduction to independent component analysis: InfoMax and FastICA algorithms. *Tutorials in Quantitative Methods for Psychology* 6, 1 (2010), 31–38.

[43] Sunghwa Lee, Younghoon Shin, Myungjong Kim, and Jiwon Seo. 2023. IR-UWB Radar-Based Contactless Silent Speech Recognition of Vowels, Consonants, Words, and Phrases. *IEEE Access* 11, 1 (2023), 144844–144859.

[44] VI Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet physics doklady* 10, 8 (1966), 707–710.

[45] Yijie Li, Juntao Zhou, Dian Ding, Yi-Chao Chen, Lili Qiu, Jiadi Yu, and Guangtao Xue. 2024. MuDiS: An Audio-independent, Wide-angle, and Leak-free Multi-directional Speaker. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. ACM, New York, NY, USA, 263–278.

[46] Somshubra Majumdar and Boris Ginsburg. 2020. MatchboxNet: 1D Time-Channel Separable Convolutional Neural Network Architecture for Speech Commands Recognition. In *Interspeech 2020*. ISCA, ISCA Cares Limited 60 Cecil Street, ISCA House, Singapore, 3356–3360. https://doi.org/10.21437/interspeech.2020-1058

[47] Geoffrey S Meltzner, James T Heaton, Yunbin Deng, Gianluca De Luca, Serge H Roy, and Joshua C Kline. 2018. Development of sEMG sensors and algorithms for silent speech recognition. *Journal of neural engineering* 15, 4 (2018), 046031.

[48] Microsoft. 2024. Windows Speech Recognition commands. https://support.microsoft.com/en-us/windows/windows-speech-recognition-commands-9d25ef36-994d-f367-a81a-a326160128c7

[49] Mumbrella. 2017. Coke No Sugar officially launches with new 'Say Yes' TV and OOH activation. https://mumbrella.com.au/coke-no-sugar-officially-launches-new-say-yes-tv-ooh-453897

[50] Kevin Murnane. 2018. Apple's Siri Is An Embarrassment. https://www.forbes.com/sites/kevinmurnane/2018/05/06/siri-is-an-embarrassment

[51] Alain Muzet. 2007. Environmental noise, sleep and health. *Sleep medicine reviews* 11, 2 (2007), 135–142.

[52] Ryohei Nakatani, Tetsuya Takiguchi, and Yasuo Ariki. 2013. Two-step correction of speech recognition errors based on n-gram and long contextual information.. In *INTERSPEECH*. International Speech Communication Association (ISCA), ISCA Cares Limited 60 Cecil Street, ISCA House, Singapore, 3747–3750.

[53] Woody norris ted talk. 2016. Woodynorris. https://www.ted.com/speakers/woody_norris

[54] Yoichi Ochiai, Takayuki Hoshi, and Ippei Suzuki. 2017. Holographic whisper: Rendering audible sound spots in three-dimensional space by focusing ultrasonic waves. In *proceedings of the 2017 CHI conference on human factors in computing systems*. ACM, New York, NY, USA, 4314–4325.

[55] Office of Justice Programs. 1993. Keeping Conversations Confidential. https://www.ojp.gov/ncjrs/virtual-library/abstracts/keeping-conversations-confidential

[56] City of Seattle. 2023. Noise Complaints. https://www.seattle.gov/police/need-help/neighborhood-issues/noise-complaints

[57] Capital One. 2024. Making the Choice Between Portrait and Landscape Display Screens. https://www.capitalone.com/cars/learn/finding-the-right-car/making-the-choice-between-portrait-and-landscape-display-screens/3091

[58] Laxmi Pandey and Ahmed Sabbir Arif. 2024. MELDER: The Design and Evaluation of a Real-time Silent Speech Recognizer for Mobile Devices. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–23.

[59] Laxmi Pandey, Khalad Hasan, and Ahmed Sabbir Arif. 2021. Acceptability of speech and silent speech input methods in private and public. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13.

[60] Fangqiang Peng. 2022. Control system and method for voice interaction between vehicle owner and pedestrian. https://worldwide.espacenet.com/patent/CN114898524A

[61] Anh Pham, Italo Dacosta, Bastien Jacot-Guillarmod, Kévin Huguenin, Taha Hajar, Florian Tramèr, Virgil Gligor, and J-P Hubaux. 2017. Privateride: A privacy-enhanced ride-hailing service. *Proceedings on Privacy Enhancing Technologies* 2017, 2 (2017), 38–56.

[62] F Joseph Pompei. 2002. *Sound from ultrasound: The parametric array as an audible sound source*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[63] Chinmay Rajguru, Daniel Blaszczak, Arash PourYazdan, Thomas J Graham, and Gianluca Memoli. 2019. AUDIOZOOM: location based sound delivery system. In *SIGGRAPH Asia 2019 Posters*. ACM, New York, NY, USA, 1–2.

[64] Jun Rekimoto. 2022. DualVoice: Speech Interaction That Discriminates between Normal and Whispered Voice Input. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 1–10.

[65] Jun Rekimoto. 2023. WESPER: Zero-shot and realtime whisper to normal voice conversion for whisper-based speech interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12.

[66] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, Vol. 2. IEEE, IEEE, Hoes Ln Piscataway, NJ, 749–752.

[67] Dongjin Seo, Hao-Yen Tang, Jose M Carmena, Jan M Rabaey, Elad Alon, Bernhard E Boser, and Michel M Maharbiz. 2015. Ultrasonic beamforming system for interrogating multiple implantable sensors. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, IEEE, Hoes Ln Piscataway, NJ, 2673–2676.

[68] Changhao Shan, Junbo Zhang, Yujun Wang, and Lei Xie. 2018. Attention-based End-to-End Models for Small-Footprint Keyword Spotting. arXiv:1803.10916 [cs.SD] https://arxiv.org/abs/1803.10916

[69] Ang Shiming and Chen Yaosen. 2005. SOUND CONTROL USING SPEAKER ARRAY. *A university wide programme, URECA or Undergraduate Research Experience on CAmpus* 1, 2 (2005), 69–72.

[70] David Shinar. 2017. Driver information processing: Attention, perception, reaction time, and comprehension. In *Traffic safety and human behavior*. Emerald Publishing Limited, Dubai Internet City P.O., 189–256.

[71] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. Muteit: Jaw motion based unvoiced command recognition using earable. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–26.

[72] Law.com Staff. 2021. Private Conversations in Public Settings: How Lawyers' 'Seemingly Innocuous Conduct' Can Lead to Confidentiality Breaches. https://www.law.com/2021/10/04/private-conversations-in-public-settings-how-lawyers-seemingly-innocuous-conduct-can-lead-to-confidentiality-breaches/?slreturn=2024111842411

[73] Andrew G Stove. 1992. Linear FMCW radar techniques. In *IEE Proceedings F (Radar and Signal Processing)*. IET, IET, Futures Place Kings Way Stevenage Hertfordshire SG1 2UA UK, 343–350.

[74] Zixiong Su, Shitao Fang, and Jun Rekimoto. 2023. Liplearner: Customizable silent speech interactions on mobile devices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–21.

[75] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018. Lip-interact: Improving mobile device interaction with silent speech commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA, 581–593.

[76] Ming Sun, Anirudh Raju, George Tucker, Sankaran Panchapagesan, Gengshen Fu, Arindam Mandal, Spyros Matsoukas, Nikko Strom, and Shiv Vitaladevuni. 2016. Max-pooling loss training of long short-term memory networks for small-footprint keyword spotting. In *2016 IEEE spoken language technology workshop (SLT)*. IEEE, IEEE, Hoes Ln Piscataway, NJ, 474–480.

[77] Xue Sun, Jie Xiong, Chao Feng, Haoyu Li, Yuli Wu, Dingyi Fang, and Xiaojiang Chen. 2024. EarSSR: Silent Speech Recognition via Earphones. *IEEE Transactions on Mobile Computing* 23, 8 (2024), 8493–9507.

[78] Eye Spy Supply. 2024. How to Protect Your Privacy During Confidential Conversations. https://blog.eyespysupply.com/2024/07/19/how-to-protect-your-privacy-during-confidential-conversations/

[79] Taobao. 2024. Sound Card. https://www.taobao.com/list/item/705139523559.html

[80] Tornex. 2023. Suggestion on Library Acoustic Performance Standards. https://www.tornex.co.kr/en/blog/acoustic-4/compare-acoustic-guideline-in-library-42

[81] IAB UK. 2021. The barriers preventing voice recognition use: speed and usefulness vs embarrassment. https://www.iabuk.com/opinions/barriers-preventing-voice-recognition-use-speed-and-usefulness-vs-embarrassment

[82] Voices. 2023. Interactive Voice Ads: Using Voice Dialogue Ads to Engage Your Audience. https://www.voices.com/blog/voice-dialogue-ads

[83] Chao Wang. 2024. *Impact of Car-Cabin Physical Environments on Driving Performance: A Multimodal Approach*. Ph. D. Dissertation. Northeastern University.

[84] Han Wang, Jiming Tang, Zhipeng Wu, and Yu Liu. 2022. A Multibeam Steerable Parametric Array Loudspeaker for Distinct Audio Content Directing. *IEEE Sensors Journal* 22, 13 (2022), 13640–13647.

[85] Lei Wang, Tao Gu, Wei Li, Haipeng Dai, Yong Zhang, Dongxiao Yu, Chenren Xu, and Daqing Zhang. 2023. Df-sense: Multi-user acoustic sensing for heartbeat monitoring with dualforming. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. ACM, New York, NY, USA, 1–13.

[86] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW based contactless respiration detection using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–20.

[87] Xue Wang, Zixiong Su, Jun Rekimoto, and Yang Zhang. 2024. Watch Your Mouth: Silent Speech Recognition with Depth Sensing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–15.

[88] Pete Warden. 2018. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. arXiv:1804.03209 [cs.CL] https://arxiv.org/abs/1804.03209

[89] Peter J Westervelt. 1963. Parametric acoustic array. *The Journal of the acoustical society of America* 35, 4 (1963), 535–537.

[90] WikiHow. 2024. How to Maintain Confidentiality. https://www.wikihow.com/Maintain-Confidentiality

[91] Philipp Wolf. 2018. An extension of the technology acceptance model tailored to wearable device technology. In *Munich Business School Working Paper*. Deutsche Nationalbibliothek, Adickesallee 1, 60322 Frankfurt am Main, 1–21.

[92] Jun Yang, Khim-Sia Tan, Woon-Seng Gan, Meng-Hwa Er, and Yong-Hong Yan. 2005. Beamwidth control in parametric acoustic array. *Japanese journal of applied physics* 44, 9R (2005), 6817.

[93] Masahide Yoneyama, Jun-ichiroh Fujimoto, Yu Kawamo, and Shoichi Sasabe. 1983. The audio spotlight: An application of nonlinear interaction of sound waves to a new type of loudspeaker design. *The Journal of the Acoustical Society of America* 73, 5 (1983), 1532–1536.

[94] Shang Zeng, Haoran Wan, Shuyu Shi, and Wei Wang. 2023. mSilent: Towards general corpus silent speech recognition using COTS mmWave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–28.

[95] Fudong Zhang, Shouxing Yuan, and Lunchuan Hu. 2016. A multi-beamforming method for parametric array. *IEICE Electronics Express* 13, 4 (2016), 20160024–20160024.

[96] Haochen Zhang, Yiyuan Wang, and Tram Thi Minh Tran. 2024. External Speech Interface: Effects of Gendered and Aged Voices on Pedestrians' Acceptance of Autonomous Vehicles in Shared Spaces. In *Adjunct Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, New York, NY, USA, 190–196.

[97] Qian Zhang, Yubin Lan, Kaiyi Guo, and Dong Wang. 2024. Lipwatch: Enabling Silent Speech Recognition on Smartwatches using Acoustic Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–29.

[98] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2021. Soundlip: Enabling word and sentence-level lip interaction for smart devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–28.

[99] Ruidong Zhang, Hao Chen, Devansh Agarwal, Richard Jin, Ke Li, François Guimbretière, and Cheng Zhang. 2023. HPSpeech: Silent Speech Interface for Commodity Headphones. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers*. ACM, New York, NY, USA, 60–65.

[100] Ruidong Zhang, Mingyang Chen, Benjamin Steeper, Yaxuan Li, Zihan Yan, Yizhuo Chen, Songyun Tao, Tuochao Chen, Hyunchul Lim, and Cheng Zhang. 2021. Speechin: A smart necklace for silent speech recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–23.

[101] Ruidong Zhang, Ke Li, Yihong Hao, Yufan Wang, Zhengnan Lai, François Guimbretière, and Cheng Zhang. 2023. EchoSpeech: Continuous Silent Speech Recognition on Minimally-obtrusive Eyewear Powered by Acoustic Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18.

[102] Yongzhao Zhang, Wei-Hsiang Huang, Chih-Yun Yang, Wen-Ping Wang, Yi-Chao Chen, Chuang-Wen You, Da-Yuan Huang, Guangtao Xue, and Jiadi Yu. 2020. Endophasia: Utilizing acoustic-based imaging for issuing contact-free silent speech commands. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–26.

[103] Yundong Zhang, Naveen Suda, Liangzhen Lai, and Vikas Chandra. 2018. Hello Edge: Keyword Spotting on Microcontrollers. arXiv:1711.07128 [cs.SD] https://arxiv.org/abs/1711.07128

[104] Qingchuan Zhao, Chaoshun Zuo, Giancarlo Pellegrino, and Li Zhiqiang. 2019. Geo-locating Drivers: A Study of Sensitive Data Leakage in Ride-Hailing Services.. In *Network and Distributed System Security Symposium (NDSS)*. CISPA, Stuhlsatzenhaus 5 66123 Saarbrücken, Germany, 1–15.

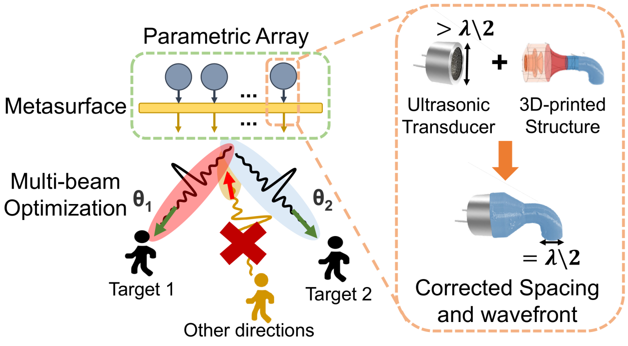


Figure 18: Prototype of the multi-directional speaker. It shows the metasurface, ultrasonic transducer array, and multi-beam optimization. The diagram also demonstrates how multiple beams can be emitted simultaneously in different directions.

[105] Jiaxin Zhong, Tao Zhuang, Ray Kirby, Mahmoud Karimi, Xiaojun Qiu, Haishan Zou, and Jing Lu. 2022. Low frequency audio sound field generated by a focusing parametric array loudspeaker. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 3098–3109.

[106] Juntao Zhou, Yijie Li, Yida Wang, Dian Ding, Yu Lu, Yi-Chao Chen, and Guangtao Xue. 2024. Visar: Projecting Virtual Sound Spots for Acoustic Augmented Reality Using Air Nonlinearity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–30.

[107] Dongsheng Zhu, Chong Han, Jian Guo, and Lijuan Sun. 2024. TWLip: Exploring Through-Wall Word-Level Lip Reading Based on Coherent SISO Radar. *IEEE Internet of Things Journal* 11, 19 (2024), 32310–32323.

A Appendix

A.1 The implementation principle of multi-directional speakers

A.1.1 Fundamental: parametric array. A parametric array is a nonlinear acoustic mechanism that generates audible sound by exploiting air nonlinearity, where two or more ultrasonic waves interact in the air to produce a difference frequency within the audible range, as described by the KZK equation [16, 62]. The received signal $r(t)$ from a transmitted signal $s(t)$ can be expressed using a summation as $r(t) = \sum_{n=1}^{\infty} \alpha_n s^n(t)$ where α_n represents the attenuation coefficient for the n-th order nonlinear term. The second-order term $s^2(t)$ is particularly important for reproducing sound from ultrasound, as higher-order terms are generally negligible.

Assume that the modulated signal expressed as $s(t) = (h(t) + 1)\cos(2\pi f_c t)$, where $h(t)$ is the low-frequency audio signal and f_c is the carrier frequency. When this signal propagates through air, the second-order nonlinear term $s^2(t)$ can be expanded as $s^2(t) = \frac{1}{2}\left(h^2(t) + 2h(t) + 1\right)\left(1 + \cos(4\pi f_c t)\right)$. Since human hearing is insensitive to the high-frequency term $\cos(4\pi f_c t)$, applying a low-pass filter leaves the low-frequency component:

$$s_{\text{audible}}(t) = \frac{\alpha_2}{2}\left[h^2(t) + 2h(t) + 1\right] = \alpha_2 \mathbf{h(t)} + \cdots$$

where α_2 is the second-order attenuation coefficient, thus enabling the parametric array to reproduce audible sound from ultrasound.

A.1.2 Realization of the multi-directionality. The prototype of the multi-directional speaker (Fig. 18) used in our system is MuDiS [45], which achieves multi-beamforming capability through spatial-division multiplexing (SDM). This method is commonly used in communication systems to transmit multiple signals simultaneously over the same frequency band but in different spatial directions. We utilize SDM to create and steer multiple independent sound beams in different directions by manipulating the phase and amplitude of the ultrasonic signals emitted from an array of transducers. Each transducer element in the phased array is carefully controlled to emit sound waves that constructively interfere in the desired directions while minimizing interference in others. The overall beamforming pattern $W(\phi)$ for a direction ϕ is given by:

$$W(\phi) = \sum_{i=1}^{n} w_i e^{j2\pi \frac{d}{\lambda}(i-1)\sin\phi}$$

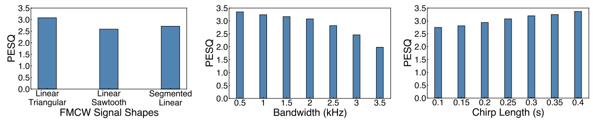
where w_i represents the complex weight applied to each transducer element, d is the spacing between elements, and λ is the wavelength of the emitted sound. By optimizing these weights for different target directions, we can project multiple beams, each carrying distinct audio content, to various spatially separated users.

The system incorporates a meticulously designed acoustic metasurface that generates a controlled wavefront and optimizes transducer spacing. The purpose of the metasurface is to redirect and focus the ultrasound emitted by each transducer, thereby producing a more precise and directional wavefront. In conjunction with the metasurface design, the multidirectional loudspeaker utilizes beam optimization algorithms to further enhance the beam-shaping process. Moreover, the system integrates a nonlinear distortion reduction mechanism to mitigate distortions arising from the nonlinearities inherent in sound wave propagation.

A.2 FMCW signal optimization

Due to the time-varying characteristics of FMCW signals, we need to carefully optimize FMCW signals. By trying different shapes of FMCW signals and selecting different FMCW signal bandwidths and chirp lengths, we will find what kind of FMCW signal will least affect the sound quality.

We compare the performance of FMCW signals with different settings for modulation. Specifically, the metric we use to evaluate the audio quality after air nonlinearity is PESQ. Fig. 19(a) shows the impact of different FMCW signal shapes, and it can be found that the linear triangular waveform results in the highest PESQ score, indicating better audio quality compared to other shapes like linear sawtooth or segmented linear. Fig. 19(b) shows the impact under different bandwidths, and it can be found that the wider the carrier bandwidth, the lower the PESQ. This is because the bandwidth of the ultrasonic speaker is limited, so the frequency response will decay rapidly within a certain range away from the center frequency, and the sound volume will decay at this time. Fig. 19(c) shows the impact under different chirp lengths. The smaller the chirp length, the lower the PESQ. This is because the frequency of the carrier changes too fast, and the vibration speed of the diaphragm of the ultrasonic array is limited, which will introduce additional noise. Considering that too narrow bandwidth and too large chirp length will affect the sensing performance, such as increasing the ambiguity of resolving with the reflected signal, we empirically select an FMCW signal with a shape of linear triangular, a bandwidth of $2kHz$, a chirp length of $0.25s$, as the carrier signal, and use it for sensing simultaneously.



(a) Impact of different FMCW signal shapes.

(b) Impact of signal bandwidth.

(c) Impact of chirp length on audio quality.

Figure 19: (a) It shows how different waveform shapes affect the perceptual evaluation of speech quality (PESQ), and linear triangular is the best. (b) The PESQ score is plotted against different bandwidth values, demonstrating that wider bandwidth leads to lower audio. (c) A plot showing the impact of different chirp lengths on PESQ, revealing that smaller chirp lengths degrade audio quality

A.3 Acoustic-based silent speech recognition principles

During silent speech, where a person articulates words without producing any audible sound, the intricate and coordinated movements of the face, lips, tongue, and even the jaw play a critical role in shaping speech sounds. These subtle articulatory gestures, though inaudible, can be effectively captured using FMCW signals. The transmitted signal interacts with the human body, and the reflected waves carry information about the movement and position of various anatomical features involved in speech production.

To extract silent speech features from the reflected signal, M^{2}Silent employs cross-correlation [86, 101]. In this context, the transmitted signal $S(t)$ is cross-correlated with the received signal $R(t)$ to produce a correlation function $C(\tau)$. This function given by:

$$C(\tau) = \int S(t)R(t+\tau)\,dt$$

reveals peaks at specific values of τ, which correspond to the time delays of the reflected signals. These time delays are indicative of the distances to various reflecting surfaces around the lip, such as the tongue. The result of the cross-correlation process is an echo frame, which is essentially a snapshot of the reflected signal characteristics at a particular moment in time. Each echo frame's element corresponds to the cross-correlation value for a specific time delay. For example, an echo frame might be represented as $[C(\tau_1), C(\tau_2), \ldots, C(\tau_n)]$ where each $C(\tau_i)$ reflects the correlation at a different delay τ_i. Multiple echo frames captured over time form an echo profile, which is critical for tracking the dynamics of facial movement corresponding to silent speech.

To reduce the impact of static noise or other consistent background reflections, we calculate a differential echo profile, which is obtained by taking the difference between consecutive echo frames: Echo Frame(t) − Echo Frame$(t-1)$. By focusing on these differences, the system can more accurately detect subtle changes in muscle movements near the lips, which are key features of silent speech.

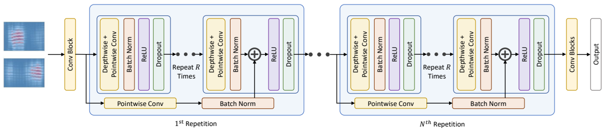


Figure 20: SilentMatch model architecture. The model uses depthwise separable convolutions, batch normalization, ReLU activation, and pointwise convolutions for silent word recognition.

A.4 Detailed description of the blind source separation algorithm

First, the mixed signals $\mathbf{F}^{mix}$ are preprocessed by centering them (subtracting the mean) and whitening them to decorrelate the signals and standardize their variances. To achieve whitening, we transform the centered mixed signals as

$$\mathbf{F}^{white} = V\mathbf{D}^{-1/2}V^{\top}\mathbf{F}^{mix},$$

where V and $\mathbf{D}$ come from the covariance matrix of the mixed signals.

We then proceed with the core of FastICA, where a demixing matrix W is determined iteratively to separate the sources by maximizing their non-Gaussianity. We start with a random weight vector $\mathbf{w}$ and update it using the rule

$$\mathbf{w}^{(new)} = \mathbb{E}\left[\mathbf{F}^{white}g(\mathbf{w}^{\top}\mathbf{F}^{white})\right] - \mathbb{E}\left[g'(\mathbf{w}^{\top}\mathbf{F}^{white})\right]\mathbf{w},$$

where g is a selected nonlinear function. After each update, we orthogonalize and normalize $\mathbf{w}^{(new)}$. This process is repeated for each source until all independent components are extracted, leading to

$$\mathbf{F}_{recovered} = W\mathbf{F}^{white},$$

which gives us the silent speech features for all users after segmentation.

A.5 Word recognition model architecture

As shown in Fig. 20, SilentMatch employs 1D time-channel separable convolutional layers, which are particularly efficient in processing temporal sequences like speech. These layers capture temporal patterns while reducing the model's computational complexity. The model is composed of 4 blocks, where each block includes a 1D convolution layer, batch normalization, ReLU activation, and a depthwise separable convolution. These layers help effectively learn the sequential dependencies in silent speech while maintaining robustness to noise and variations. Furthermore, we reduced the stride in the convolutional layers to better capture fine-grained silent word features and reduced the kernel size to allow the model to capture more detailed silent speech features. For training, we utilize the cross-entropy loss function, and the model is optimized using SGD with momentum.

SilentMatch outputs a prediction for a word based on the input features. If the confidence level for all possible word predictions is low, indicating uncertainty in the prediction, the model will output a blank instead of forcing an incorrect word prediction. This mechanism ensures that only confidently recognized words are considered, reducing errors in silent word recognition.