Robust Dual-Modal Speech Keyword Spotting for XR Headsets

Zhuojiang Cai 🝺, Yuhan Ma 🝺, and Feng Lu 🝺, Senior Member, IEEE



Fig. 1: Our dual-modal system significantly extends the available scenarios for speech interaction with XR headsets. (a) We propose a vocal-echoic dual-modal keyword spotting system. (b) In typical environments, such as working or entertaining indoors, dual-modal system performs accurately and can block out the speech of speakers nearby. (c) In noisy environments, such as on the street or in the subway, our dual-modal method outperforms their single-modal counterparts. (d) When unable or unwilling to make a voice, for example, when others are working or sleeping nearby, our system continues to function effectively.

Abstract—While speech interaction finds widespread utility within the Extended Reality (XR) domain, conventional vocal speech keyword spotting systems continue to grapple with formidable challenges, including suboptimal performance in noisy environments, impracticality in situations requiring silence, and susceptibility to inadvertent activations when others speak nearby. These challenges, however, can potentially be surmounted through the cost-effective fusion of voice and lip movement information. Consequently, we propose a novel vocal-echoic dual-modal keyword spotting system designed for XR headsets. We devise two different modal fusion approches and conduct experiments to test the system's performance across diverse scenarios. The results show that our dual-modal system not only consistently outperforms its single-modal counterparts, demonstrating higher precision in both typical and noisy environments, but also excels in accurately identifying silent utterances. Furthermore, we have successfully applied the system in real-time demonstrations, achieving promising results. The code is available at https://github.com/caizhuojiang/VE-KWS.

Index Terms— Speech interaction, extended reality, keyword spotting, multimodal interaction.

1 INTRODUCTION

In recent years, Extended Reality (XR) headsets, including Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR) headsets, have gained widespread attention. These devices, serving as bridges between the virtual and real worlds, have the potential to profoundly reshape how people live, work, and entertain themselves. With the increasing popularity of MR and AR headsets like the Apple Vision Pro and Microsoft HoloLens, XR headsets are finding utility in an ever-expanding array of scenarios.

Efforts to develop natural and effective user interaction methods with XR headsets have been a critical research focus. A growing number of headsets, including the Apple Vision Pro, have shown a preference for intuitive interaction methods such as speech, gestures, and gaze. These methods, in comparison to tools like controllers, are more readily accepted and user-friendly. Among them, speech interaction can be used for menu selection, locomotion control, and combined with gaze for object movement, etc. However, the extensive use of speech interaction has been hindered by limitations in speech recognition methods, including Automatic Speech Recognition (ASR) and Speech Keyword Spotting (KWS), which lack robustness across various real-world environments.

- Zhuojiang Cai, Yuhan Ma, and Feng Lu are with State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China. E-mail: {caizhuojiang | raphael.mayuhan | lufeng}@buaa.edu.cn.
- Zhuojiang Cai and Yuhan Ma contribute equally to this work. Feng Lu is the corresponding author.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxx/TVCG.201x.xxxxxx

Conventional vocal keyword spotting methods encounter various challenging usage scenarios. Specifically, we have identified three common and demanding situations: 1) In noisy environments such as bustling streets, crowded malls, or noisy public transport, keyword spotting accuracy significantly degrades. 2) Users often cannot vocalize when others are working or resting nearby, or due to privacy concerns and social awkwardness. In such cases, vocal keyword spotting fails entirely. 3) Vocal speech keyword spotting systems are susceptible to interference and even false triggering when others are speaking nearby.

These challenges highlight the limitations of traditional vocal keyword spotting approaches and emphasize the need for more robust and versatile solutions. Addressing the second challenge, significant progress has been made in the field of silent speech interfaces. Various types of data, including facial imagery [11,52], ultrasonic imaging [22], EMG [17, 25, 48], motion sensing [37], strain sensing [24], and other sensing forms have been investigated for recognizing silent speech. Furthermore, some research has leveraged the transmission and reception of ultrasonic waves to detect mouth movements [13, 55, 56], offering contactless and cost-effective solutions. EchoSpeech [55], for instance, employed Frequency-Modulated Continuous Wave (FMCW) and demonstrated impressive silent speech spotting capabilities using off-the-shelf speakers and microphones mounted on glasses frames. However, this work did not explore the use of low-frequency vocal speech information from microphone audio, which could extend its application to a broader range of vocal speech recognition scenarios, thereby addressing the remaining two challenges.

Recognizing that vocal speech and lip movement can contribute from different angles to the comprehension of a speaker's speech, and drawing inspiration from related research in audio-visual speech recognition [9, 34, 42, 57] that supports this idea, we posit that vocal speech information and mouth movement information obtained through ultrasonic echo could similarly offer diverse perspectives on a speaker's speech. The integration of these two modalities may provide a wealth of prior knowledge for speech recognition, potentially leading to improved performance.

Therefore, this paper introduces a novel vocal-echoic dual-modal keyword spotting method for XR headsets. By combining vocal speech and ultrasonic echo features, it achieves robust keyword spotting in various scenarios. We designed and implemented this dual-modal keyword spotting system on Microsoft HoloLens 2 with custom hardware. We devised two different modal fusion approaches and verified their effectiveness through experiments. Furthermore, we compared the dual-modal system with single-modal systems in low-noise environments, diverse noisy environments, and scenarios with interference from other speakers. The results demonstrate that our proposed dual-modal system consistently outperforms its single-modal counterparts in the majority of scenarios, without compromising silent speech spotting performance. Finally, we applied the system in practical examples, achieving promising results. Our work significantly broadens the scope of speech keyword spotting applications.

In summary, our contributions can be outlined as follows:

- We propose a vocal-echoic dual-modal speech keyword spotting (KWS) system, enabling robust speech recognition in a broader range of scenarios.
- We conducted an ablation study to design a lighter CNN model for the echoic modal KWS, reducing its demand for computing resources in the headset.
- We conducted experiments to assess the system's performance in various noisy environments, situations with nearby speakers, and silent scenarios, demonstrating that our dual-modal system outperforms traditional vocal systems in all of these scenarios.

2 RELATED WORK

2.1 Vocal Speech Keyword Spotting

Keyword Spotting (KWS) is the task of detecting a predefined set of keywords from an audio stream. Compared to Automatic Speech Recognition (ASR), KWS can be deployed on edge devices with low computational requirements and without the need for cloud connectivity, eliminating privacy concerns. Therefore, it finds extensive applications in various domains such as user interface interactions, smart home control, and command triggers.

Typically, KWS refers to the detection of vocal speech. However, in this paper, we also employ a silent speech keyword spotting using echoic modality. To differentiate, we refer to the conventional KWS as vocal speech keyword spotting in this context.

ASR-based KWS. Some studies utilize ASR systems to convert speech signals into text and then identify keywords through text matching techniques [12, 31, 40]. While this approach eliminates the need for specialized training of predefined keywords, offering greater flexibility, it inherits the drawbacks of ASR, including computational demands and privacy concerns.

HMM-based KWS. The KWS system based on Hidden Markov Models (HMM) was proposed three decades ago [38, 39, 51]. In this approach, speech samples of each keyword are used to train the corresponding HMM for that keyword, and non-keyword (filler) speech segments are used to train a filler HMM. At runtime, the input audio stream is matched against these HMM models. The Viterbi algorithm is commonly employed to find the most likely state sequence. If the matching probability exceeds a predefined threshold, the system identifies the input as containing the keyword. Although these systems perform well, executing multiple HMM model matches with the Viterbi algorithm still requires significant computational power [4, 29].

Deep KWS. In the past decade, the rapid advancement of deep learning has led to extensive research and application of deep KWS approches [29]. These advancements have resulted in reduced computational complexity and improved performance of KWS systems. Common architectures, such as fully-connected networks [4] and Recurrent Neural Networks (RNNs) [23, 43, 58], have been studied and proven effective in KWS systems. Inspired by computer vision research, KWS

based on Convolutional Neural Networks (CNNs) [6, 19, 41, 44, 53] has garnered significant attention due to its straightforward architecture and ease of tuning [44]. Notably, Tang and Lin [44] applied the concept of residual learning to KWS, creating high-performance models with a small footprint. Subsequently, TC-ResNet [6] replaced the original convolutional modules in residual blocks with temporal convolutions, and DC-ResNet [53] introduced depthwise separable convolutions. These innovations reduced the parameter count and computational complexity. Kim et al. [19] introduced a novel network architecture called Broadcast Residual Learning, which consistently outperforms previous models with an equivalent parameter count. However, despite demonstrating excellent performance on public datasets, these methods experience significant performance degradation in noisy environments [7, 30, 35]. This issue will be thoroughly investigated in this paper.

2.2 Silent Speech Interface on Wearables

Silent Speech Interface (SSI) can be viewed as a generalized form of speech recognition. It involves capturing non-vocal information using various sensors to recognize speech utterances from users.

Contacting SSI. In wearable devices, SSIs implemented through diverse sensor technologies have been widely researched. Some approaches place magnetometers [1,3,15] or capacitive sensors [20,28] inside the mouth to capture movements of the mouth and tongue, reconstructing speech utterances. Others employ sensors attached closely to the skin, recognizing speech through forms of information like ultrasound imaging [22], Electromyography (EMG) [17,25,48], motion sensors [37], or stress sensors [24]. However, these methods require skin contact or even intrusion into the mouth, which may be uncomfortable for users [55].

Contact-free SSI. Contact-free SSIs offer a more comfortable user experience, leading to increased research interest. Some efforts attempted to install cameras on headphones [5] or neck-mounted devices [21, 54]; however, these methods faced challenges related to high power consumption and privacy risks. Additionally, some works utilize active acoustic methods, employing microphones and speakers on smartphones [13], VR headsets [56], and glass-frames [55] to achieve low-power, high-performance silent speech keyword spotting. Among them, CELIP [56] requires a setup directly facing the user's mouth, including a pair of relatively large microphones and speakers, whereas EchoSpeech [55] integrates compact components under the lower frame of glasses, ensuring a less obtrusive design.

Although these contact-free methods achieve silent speech recognition with low power consumption and high performance, their potential to integrate with vocal keyword spotting systems for enhanced performance and broader application scenarios remains unexplored.

2.3 Speech Interaction in XR

Speech interaction allows users to control devices in XR through verbal commands, either independently or in collaboration with other interaction methods like controller, gesture, and gaze. This approach provides a more flexible and user-friendly way to operate devices in XR.

Hand-free Interaction. The use of speech commands for hands-free interaction is a common practice in XR [14, 16, 36]. For instance, voice commands are utilized to control locomotion in VR without the need for hand gestures [16]. In dental implant surgeries where hands might be occupied, speech commands can replace menu clicks in AR [36]. Studies suggest that voice-based interactions can be as efficient as gesture-based interactions [26] and free up hands for other tasks [32].

Multimodal Interaction. Other studies have explored interactive methods combining speech commands with gesture and gaze. For example, using gestures to select a target and then executing operations through speech commands [33] is considered more efficient and accurate than relying solely on gestures [26]. Another approach involves using gaze and speech to control the movement of objects [10, 18]. Wang et al. [49] investigated a multimodal interaction method in AR that combines speech, gaze and gesture and demonstrated its superior efficiency compared to single-modal and dual-modal methods.



Fig. 2: Overview of vocal-echoic dual-modal KWS system for XR headset. (left) Hardware diagram of experimental equipment. The speakers and microphones are mounted on the XR headset, connected to an ESP32, which sends the audio to a PC over the network. The PC is used for algorithm implementation and experiments, and it sends the detected keywords back to the application on the headset. (right) Algorithm flowchart. The audio is separately filtered and input into the vocal and echoic modal KWS pipelines. The predicted vectors obtained from these pipelines are then fed into the fusion module to generate the keyword output.

3 DUAL-MODAL KEYWORD SPOTTING SYSTEM

3.1 System Overview

Our dual-modal keyword spotting (KWS) system represents an elegant enhancement of its commonly used single-modal vocal counterpart. It maintains audio streams as the sole input and predicted keywords as the output.

Unlike traditional systems that directly use input audio for vocal KWS, our system separates audio into vocal and echoic modalities using bandpass filters. The echoic audio originates from ultrasonic waves emitted by speakers and reflected off the user's skin near the mouth. These two segments of audio are then separately processed by the vocal and echoic modal KWS pipelines before their predictions are fused into a single output.

The difference in frequency ranges between vocal and ultrasonic echoic audio makes this system feasible. Research has shown that the fundamental frequency (F0) of vocal speech typically falls within the range of approximately 100-240Hz. Even considering the harmonic information that may be present in vocal speech, the upper limit of the frequency bands used in conventional speech recognition Mel-filter banks is usually around 5000Hz. In contrast, the ultrasonic waves used in our system operate at frequencies above 17kHz. Hence, both modalities of audio can be concurrently captured using the same microphone, providing representations of two modalities for the same speech.

Both modal KWS pipelines in the system employ lightweight deep learning approaches. For vocal KWS, many methods have already achieved excellent results in typical low-noise scenarios. In this paper, we directly utilize one of the state-of-the-art convolutional neural networks (CNNs) for vocal KWS, which will be discussed in Sec. 3.3. For echoic KWS, EchoSpeech [55] uses a ResNet18 as the backbone network. We significantly reduced the network's parameters with negligible performance degradation through an ablation study, enabling it to run with lower power consumption and latency. This will be discussed in Sec. 3.4.

The predictions from the two KWS pipelines are fused in the final stage of the system. We propose two fusion methods: reliabilitybased fusion and MLP-based fusion. Both have been experimentally shown to achieve higher accuracy than single-modal results. This demonstrates that our fusion methods can comprehensively consider different dimensions of information from the two modalities for the same speech, resulting in more accurate keyword spotting. These two fusion methods will be introduced in Sec. 3.5.

3.2 Hardware Implementation

Considering that the performance of the echoic modality is highly dependent on the positions of the speakers and microphones, it is worth



Fig. 3: Hardware Setup. (a-b) Front view and bottom view of our implementation with HoloLens headset. (c) ESP32 Add-on board.

noting that the positions of the components integrated into commercial XR headsets do not align optimally with the requirements of our system.

Therefore, we employed a microcontroller board and a customdesigned add-on board to connect two pairs of small off-the-shelf speakers and microphones, which were mounted on Microsoft HoloLens 2 to implement the hardware for this system, as shown in Fig. 3 (a-b). This positional configuration has demonstrated strong performance in glasses-frames [55] for echoic modal KWS.

The microphones and speakers are both mounted on the lower edge of the lenses of HoloLens 2. Two speakers are along the lower edge of the right lens, while two microphones are along that of the left lens. The direction of the speakers and microphone holes is oriented downward. The distance between the speaker and the microphone closer to the nose is 7.2 cm, while the distance between the two speakers and between the two microphones is 1.8 cm. Additionally, The average vertical distance from the microphone plane to the horizontal midline of the mouth is 3.9 cm, measured across 15 participants wearing the device.

During the keyword spotting process, user utterances are captured by the microphones. Simultaneously, the speakers emit continuous ultrasonic waves, with a portion being reflected off the user's face and mouth, subsequently captured by the microphones as echoes. Hence, the signals received by the microphones can be separated into two components: vocal modality and echoic modality, enabling dual-modal keyword spotting and fusion.

In the experimental setup, an ESP32 development board is employed to control speaker playback, receive microphone signals, and transmit them to a computer or HoloLens for subsequent detection and fusion. The development board, speakers, and microphones are all common products purchased online. The development board used is the ESP32-S3-DevKitC-1-N8R8, the speakers are OWR-05049T-38D, and the



Fig. 4: (a) The frequency-time diagram of the FMCW signals in the two frequency bands. (b-c) Original and differential Echo Profile.

microphones are ICS-43434. Furthermore, an add-on board has been designed to interface the modules with the ESP32 and decode audio, as shown in Fig. 3 (c).

3.3 Vocal Modal KWS

Keyword Spotting (KWS) in the vocal modality is a relatively mature technology, with a substantial amount of state-of-the-art work currently utilizing deep learning approaches [2, 6, 19, 44, 46]. These methods involve converting audio into Mel-frequency cepstral coefficients (MFCCs) and feeding the MFCCs into convolutional neural networks (CNNs) or other neural networks to produce keyword spotting results, which is the methodology our system adopts.

In our dual-modal KWS system, audio is first processed with a low-pass filter set to a 10 kHz cutoff frequency to isolate vocal audio. This audio is then subjected to a series of transformations, including pre-emphasis, framing, windowing, fast Fourier transform (FFT), Melfrequency warping, logarithmic scaling, and discrete cosine transform, ultimately resulting in MFCCs features [8, 29]. These features are fed into a broadcast residual based CNN architecture [19] to generate prediction vectors.

Differing from single-modal vocal KWS systems, where the *argmax* in the prediction vector typically serves as the system output, we combine this vector with the one from the echoic modality KWS pipeline and input both vectors into a fusion module to obtain a unique prediction result.

3.4 Echoic Modal KWS

The echoic modal KWS pipeline utilizes an active acoustic approach. Speakers installed at the bottom of the XR headset emit Frequency Modulated Continuous Waves (FMCW) in two frequency bands. Simultaneously, microphones, also positioned at the bottom, receive audio signals reflected off the skin near the mouth (echoes). By computing the cross-correlation between the echoes and the transmitted signals in each FMCW frame, features about mouth movements can be obtained, which can be learned by a Convolutional Neural Network (CNN) to achieve keyword spotting.

Why FMCW? Different active acoustic methods, including Doppler effect, CIR, and FMCW, have been employed in prior works for silent speech interfaces. All of these methods have the potential to replace our echoic modality pipeline, as our fusion approach solely takes the keyword classification probability vectors as inputs. Among these methods, a previous study demonstrated the advantages of FMCW in capturing facial movements [27]. Therefore, we have chosen this method to ensure optimal results after fusion.

Transmitted signal. Our transmitted signals are chirps in FMCW, where the frequency f linearly increases over time t within each chirp. This can be represented as $f(t) = f_l + (f_h - f_l) \times t/T$, where f_l and f_h represent the lower and upper bounds of the frequency, and T represents the chirp period, as illustrated in Fig. 4 (a). In our system, two microphones simultaneously emit FMCW signals in two frequency bands: 17-20 kHz and 20.5-23.5 kHz, with a period of 12 ms. This frequency range is considered ultrasonic and is compatible with the sampling rates of commercial microphones and speakers, which operate at 48 kHz.

Cross-correlation based FMCW. The cross-correlation based FMCW method [47] is employed to process the echoes received by the

microphone. By computing the cross-correlation between the echoes and the transmitted signals, the correlation at different sampling offsets within one chirp can be obtained. The sample shift is proportional to the distance from the microphone and speaker to the reflecting medium. Therefore, the cross-correlation reflects the magnitude of reflection at different distances, allowing for the resolution of 0.357 cm when the sampling rate is 48 kHz. This level of resolution enables the detection of positional changes of the skin near the user's mouth relative to the XR headset during speech.

Echo Profile. Calculating the cross-correlation for each consecutive chirp of the received signal produces a correlation graph with time on the horizontal axis and sample shift on the vertical axis, refer to as the Echo Profile [27]. The differential Echo Profile, obtained by differencing the Echo Profile along the time axis, reveals the temporal movement characteristics of the mouth (see Fig. 4 (b-c)).

CNN Model. In echoic modal KWS, the differential Echo Profile serves as the input to the CNN network to predict keywords in the audio. Previous study [55] utilized ResNet-18 as the backbone network, achieving good performance. However, ResNet-18 has significantly more parameters compared to the network used in vocal modality KWS, leading to additional computational demands. Therefore, we used ResNet-18 as the baseline and improved the architecture of the echoic modality network through ablation study. We achieved this by reducing the network width and replacing regular convolutional modules with depthwise separable convolutions, resulting in multiple models with different parameter counts. The ablation study is detailed in Sec. 4.3. We selected ResNet-18-1/4-DS as the network for our echoic KWS pipeline because it strikes a balance between performance and parameter count.

3.5 Fusion Strategies

The fusion strategy comprises two objectives: on one hand, fully leveraging the distinct representations of user speech provided by both modalities to enhance prediction accuracy; on the other hand, remaining unaffected by unreliable information from one modality and utilizing trustworthy information from the other modality.

To achieve these goals, we explore two fusion methods: reliabilitybased fusion and MLP-based fusion. The former employs manually crafted features, calculating reliability indicators of prediction vectors from both modalities' pipeline outputs. These indicators are then adaptively used to determine fusion strategies and perform fusion operations. The latter employs a neural network approach, utilizing a Multi-Layer Perceptron model to learn the relationship between prediction vectors from both modalities and the fusion results.

3.5.1 Reliability-based Fusion

From the two objectives, as we aim to prevent the influence of unreliable information from individual modalities (for instance, the impact of vocal modality predicting "silence" during silent speech recognition), a natural approach is to assess the reliability of predictions from each individual modality. Therefore, we devised a reliability-based fusion strategy, adaptively fusing modalities based on reliability indicators.

Reliability indicator. We utilized a reliability index based on N-best log-likelihood difference and N-best log-likelihood dispersion. This fusion method, as demonstrated by Potamianos et al. [34], has shown effectiveness in the realm of audio-video integration.

The first indicator is used to measure the class discrimination ability of each modality, while the second indicator supplements the additional N-best class likelihood ratios missing in the first indicator. The amalgamation of these two indicators allows for a more rational evaluation of the modality's reliability. The definitions of the two indicators are as follows:

N-best Log-Likelihood Difference:

$$L_{m,t} = \frac{1}{N-1} \sum_{n=2}^{N} \log \frac{P(\mathbf{o}_{m,t}|c_{m,t,1})}{P(\mathbf{o}_{m,t}|c_{m,t,n})}$$
(1)

N-best Log-Likelihood Dispersion:

$$D_{m,t} = \frac{2}{N(N-1)} \sum_{n=1}^{N} \sum_{n'=n+1}^{N} \log \frac{P(\mathbf{o}_{m,t}|c_{m,t,n})}{P(\mathbf{o}_{m,t}|c_{m,t,n'})}.$$
 (2)

Here, $P(o_{m,t}|c_{m,t,n})$ denotes the likelihood of observing the result $o_{m,t}$ given the class $c_{m,t,n}$. In these equations, *m* and *t* represent the modality and time respectively, and *n* indicates the class ranking.

Fusion process. During the fusion process, reliability indicators are first employed to separately determine whether the prediction vectors from the two modalities are utilized. When both modalities are utilized, the reliability indicators are then used to calculate the fusion exponent.

To determine the utilization of a modality, a threshold mechanism $[t_{l,m,t}, t_{d,m,t}]$ has been implemented. A modality is considered reliable only if its reliability indicators meet specific threshold criteria. If one modality is deemed unreliable, we rely solely on the modality considered reliable to ensure the accuracy of the fusion results. This approach prevents interference from the unreliable modality. The boolean variable $R_{m,t}$ indicates whether the modalities exceed their respective thresholds, determining their reliability and whether their data will be included in the subsequent fusion process. The calculation of $R_{m,t}$ is defined as:

$$R_{m,t} = (L_{m,t} > t_{l,m,t}) \land (D_{m,t} > t_{d,m,t}).$$
(3)

In situations where both modalities are considered reliable, the information from both modalities should be fully utilized to obtain a more accurate result. In this case, we will reapply the reliability indicators to determine the fusion exponent $\lambda_{v,t}$:

$$\lambda_{v,t} = \frac{1}{1 + \exp\left(-\sum_{i=1}^{4} w_i a_{i,t} d_{i,t}\right)}.$$
 (4)

Here, $d_t = [L_{v,t}, D_{v,t}, L_{e,t}, D_{e,t}]$ corresponds to the four reliability indicators of two modalities. w_i assigns weights to the reliability indicators to ensure that the distinct reliability indicators of different modalities are appropriately aligned during the fusion process. It is important to emphasize that these weights are solely related to the vocal and echoic KWS models involved in the fusion and remains unaffected by changes in the data.

Furthermore, $a_{i,t}$ signifies the reliability adjustment for two classes: "silence" and "unknown". Considering their auxiliary roles in the classification, when one modality provides "silence" or "unknown", it is imperative to take into full consideration the information from the other modality, thus necessitating a reduction in their reliability within a finite range. Differently, "silence" provided by the vocal modality should be accorded special attention. Typically, there is virtually no possibility of vocalization without any facial muscle movement. The value of $a_{i,t}$ will only change to a specific value when there is at least one modality providing "silence" or "unknown"; in all other cases, it remains a four-dimensional vector consisting of ones.

Subsequently, we weight the decisions of the two single-modal classifiers and utilize their log-likelihoods for linear combination, thereby obtaining the fusion results. Overall, our fusion process can be described by:

$$\begin{cases}
P(\mathbf{o}_{f,t}|c) = P(\mathbf{o}_{v,t}|c)^{\lambda_{v,t}}P(\mathbf{o}_{e,t}|c)^{(1-\lambda_{v,t})}, & R_{v,t} \land R_{e,t} \\
P(\mathbf{o}_{f,t}|c) = P(\mathbf{o}_{v,t}|c), & R_{v,t} \land \neg R_{e,t} \\
P(\mathbf{o}_{f,t}|c) = P(\mathbf{o}_{e,t}|c), & \neg R_{v,t} \land R_{e,t} \\
No credible result, & \neg R_{v,t} \land \neg R_{e,t}.
\end{cases}$$
(5)

Determination of Parameters. This fusion strategy initially requires the determination of eight parameters, four for setting thresholds and the other four for calculating exponents. This task employs a genetic algorithm to maximize the objective function that reflects the accuracy of the fusion results under different parameters. The remarkable global search capability of the genetic algorithm enables it to achieve satisfactory optimization effects under complex objective functions. In order to obtain a parameter vector of as high-quality as possible, we used the Latin Hypercube Sampling (LHS) initialization method to maximize coverage of the available parameter space for better search capability. The initial solution obtained with fewer iterations on a small dataset is used to replace the optimal solution in the initial population to cope with the high time cost of the genetic algorithm and its dependence on the initial population.

In addition to determining the aforementioned eight parameters, it is imperative to ascertain the target values for parameter $a_{i,t}$ when both modalities yield "silence" and "unknown" results separately. We need to determine four parameters, denoted as $a_{v,s}$, $a_{v,u}$, $a_{e,s}$, and $a_{e,u}$, corresponding to the modifications when each of the two modalities provides "silence" and "unknown" outcomes.Once the above eight parameters are established, we can employ a grid search to swiftly determine the values of these four parameters.

3.5.2 MLP-based Fusion

The MLP-based fusion method abandons manually crafted features and employs a straightforward Multi-Layer Perceptron (MLP) model to generate fusion results. It consists of an input layer, a hidden layer, and an output layer, with each layer fully connected to the next. The input vector is formed by concatenating the output vectors obtained from the vocal and echoic modal pipelines. This approch can be represented as follows:

$$P(\mathbf{o}_{f,t}|c) = MLP([P(\mathbf{o}_{v,t}|c), P(\mathbf{o}_{e,t}|c)]).$$
(6)

Here, $P(\mathbf{o}_{f,t}|c)$ represents the fusion result at time *t* given class *c*, $P(\mathbf{o}_{v,t}|c)$ and $P(\mathbf{o}_{e,t}|c)$ represent the prediction vectors for vocal and echoic modalities, respectively. The square brackets indicate the concatenation of the two output vectors.

We utilized the self-recorded dataset introduced in Sec. 4.1 to train the model. During training, this dataset was rigorously partitioned, with a strict demarcation between the data used for training and the data employed in subsequent experiments. Multiple data augmentation techniques were employed during training. For each data instance, four possible processing methods were applied: preserving clarity (no processing), introducing environmental noise interference, introducing vocal noise interference, and discarding a portion of vocal data. The environmental noise used in processing was extracted from the DE-MAND [45] noise dataset, as detailed in Sec. 4.4. This noise data was then scaled to random signal-to-noise ratios before being combined with the training data. Additionally, vocal noise data was sourced from the Google Speech Commands [50] dataset, as explained in Sec. 4.6. This vocal noise data was added to the training data after being multiplied by a fixed coefficient. Among these four processing methods, only one was randomly selected. Subsequently, the vector was multiplied by a random factor ranging between 0.95 and 1.05 to simulate random noise interference. Furthermore, the cross-entropy loss function and the Adam optimizer were employed during the training process.

In application, the system concatenates the prediction vectors from both the vocal and echoic modalities, and feed the resulting concatenated vector into the trained multi-layer perceptron. Subsequently, we select the class with the highest probability from the resulting prediction vector as the system's output.

4 EXPERIMENTS

4.1 Data

We recruited 15 participants (2 female and 13 males, from 20 to 27 years old) wearing our equipment and reading specific keywords to record audio data. Each participant read 30 repetitions of 10 comand words, and 5 repetitions of another 25 auxiliary words labeled as "unknown" to help distinguish unrecognized words. Additionally, each participant also wore the device to record data in which they did not speak and kept their mouth still, labeled as "silence". The data has a sampling rate of 48 kHz, allowing for the preservation of audio information with a maximum frequency of 24 kHz. As a result, the recorded data contains information from both vocal and echoic modalities.

The vocabulary used in the experiments is derived from the Google Speech Commands [50] dataset, which is the most commonly used open-source dataset for vocal KWS task. Therefore, it was natural for us to migrate the vocabulary from this dataset to our dual-modal KWS experiments. This allows our vocal modality model to be trained on a combination of this dataset, which has a large volume of data, and our proprietary dataset, maximizing its performance and robustness. This approach ensures that the comparative experimental results between single-modal and dual-modal performance are reliable.

4.2 Evaluation Metric

Word Error Rate (WER). The Word Error Rate (WER) is a metric used to evaluate the performance of speech recognition systems or natural language processing systems. It measures the difference between two texts, i.e., how many words in the predicted text are incorrect, missing, or redundant. The WER can be calculated as:

$$WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C}$$
(7)

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words, and N is the total word count in the text.

4.3 Experiment 1: Echoic Model

In this experiment, we conducted an ablation study on the echoic modality deep learning model using our proprietary dataset. The primary objective was to investigate the trade-off between model complexity (parameter count) and accuracy, with the ultimate goal of achieving a more streamlined and lightweight model.

4.3.1 Experiment Setup

We commenced our experiments with ResNet-18, a model that EchoSpeech has previously validated for its strong performance in echoic modal KWS.

Our initial approach involved reducing the number of channels within the convolutional layers, referred to as width of the network. This adjustment was motivated by the fact that ResNet-18 was originally designed for tasks such as image classification and other computer vision applications, where visual data typically carries a richer information load compared to audio data. Consequently, reducing the network's width was a deliberate strategy aimed at decreasing the model's parameter count and mitigating the risk of overfitting. In our study, we explored width reduction in comparison to the original ResNet-18, as well as versions with widths reduced to 1/2, 1/4, and 1/8 of the original width. We assessed their respective performance within the echoic KWS pipeline.

Furthermore, we employed depthwise separable convolutions to replace convolution modules in the model. Depthwise separable convolution modules break down standard convolutions into depthwise convolution and pointwise convolution, significantly reducing the number of parameters and computations while maintaining similar performance. We compared the accuracy and parameter count differences when using depthwise separable convolution modules with different widths.

Our experiments were conducted on approximately 1800 speech samples from five participants, with a data split of 80% for training and 20% for testing. Random seeds were used to control data splitting and data augmentation parameters. For each model variant, we systematically adjusted the random seeds and conducted ten rounds of training and testing to obtain a reliable average accuracy. Furthermore, we employed *torchstat* to analyze the model's parameters and computational complexity.

In the training process, we utilized the SGD optimizer with a learning rate following warm-up strategy, starting from 0 and linearly increasing to 0.1 over the first 50 epochs. Subsequently, the training continues for 1000 epochs with cosine learning rate decay. Additionally, we applied data augmentation including random noise, random padding, and overlaying background noise data.

4.3.2 Results and Discussion

The results are presented in Tab. 1, illustrating eight different model configurations resulting from combinations of four distinct network widths and the inclusion of depth-wise separable convolutions (DS). This table provides insights into the average accuracy, parameter count, and computational complexity of these models.

In the model naming convention, the fraction following *ResNet-18* denotes the reduction in model width relative to the original model, and the suffix *DS* indicates the utilization of depth-wise separable convolutions.

Notably, we observed that transitioning from ResNet-18 to ResNet-18-1/4-DS resulted in only a marginal 0.91% increase in average WER, while substantially reducing the parameter count by over 100 times. This highlights the potential for our models to be efficiently deployed on XR headsets.

Table 1: Comparison of	different models'	performance,	including Word
Error Rate, Parameters,	and Multiply-Add	Operations.	

Model	WER	Params	MAdd
ResNet-18	$5.06\% \pm 0.650$	11.22M	1.84G
ResNet-18-1/2	$5.47\% \pm 0.466$	2.820M	468.4M
ResNet-18-1/4	$5.81\% \pm 0.639$	712.6K	121.11M
ResNet-18-1/8	$7.69\% \pm 0.528$	182.0K	32.28M
ResNet-18-DS	$5.33\% \pm 0.539$	1.484M	264.1M
ResNet-18-1/2-DS	$5.56\% \pm 0.714$	394.0K	79.99M
ResNet-18-1/4-DS	$5.97\% \pm 0.573$	109.9K	24.74M
ResNet-18-1/8-DS	$8.00\% \pm 1.260$	33.22K	8.93M

4.4 Experiment 2: Noisy Environment

In this section, we examined the performance of two fusion strategies across different environments and signal-to-noise ratios (SNR). The research findings indicate that, in all experimental conditions, the performance of the Keyword Spotting (KWS) system using Vocal-Echoic dual-modal fusion surpasses that of its single-modal counterparts.

4.4.1 Experiment Setup

In this experiment, we simulated diverse scenarios by superimposing data with noise of varying intensities corresponding to specific scenes. We compared the average performance of different modalities within these scenarios to evaluate their effectiveness.

The signal-to-noise ratio (SNR) was employed to straightforwardly measure the strength of noise, defined as the ratio of signal power P_s to noise power P_n . Decibels (dB) were utilized as the unit of measurement for SNR, as shown in the following formula:

$$SNR(dB) = 10\log_{10}P_s/P_n \tag{8}$$

We investigated the performance of the fusion modality in environments with SNR ranging from -10dB to 10dB. The noise signals were multiplied by coefficients, which were determined based on the average power of the data and the target SNR, and then added to the data to simulate various intensity levels of noisy environments. The data were then subjected to vocal and echoic modality processing steps, as previously described, to generate predictive results. Subsequently, the fusion modality produced fused results. We compared the word error rates of these results to assess the performance of different modalities in these noisy environments.

4.4.2 Results and Discussion

In Fig. 6, we present the average results of two single modalities: vocal and echoic, and two dual-modal fusion strategies: reliability-based fusion (RB fusion), and MLP-based fusion (MLP fusion) across all scenarios. The standard deviations of these data are also depicted with shaded regions of the same color on the graph. Experimental results show that both MLP fusion and RB fusion outperform the two individual modalities.



Fig. 5: Frequency distribution of Meeting (left) and Metro (right). The frequency distribution of noise varies across different scenarios, with some noise having a significant presence in high frequencies, while another portion is primarily concentrated in lower frequencies.



Fig. 6: Comparison of average Word Error Rates (WER) between singlemodal and dual-modal KWS systems in all noise scenarios. Our dualmodal systems (RB Fusion and MLP Fusion) achieve lower WER than single-modal systems (Echoic and Vocal) across all SNRs. At the strongest noise level (SNR=-10.0), MLP fusion reduces WER by 15.68% and 16.57% compared to vocal and echoic systems.

To provide a more detailed assessment of the system's performance across different modalities, we illustrate the experimental results of RB fusion in six distinct scenarios in Fig. 7 (a). RB fusion exhibits slightly higher average word error rates, making it a more challenging test of fusion strategy performance. The results for each scenario are calculated as the average of sub-scenario experimental outcomes. In the first three scenarios, such as the Domestic setting with pronounced high-frequency noise interference, the vocal modality exhibits superior performance over the echoic modality. Conversely, in the latter three scenarios, such as the Public setting where noise primarily manifests as low-frequency disturbances, the echoic modality demonstrates a performance advantage over the vocal modality. Regardless of the individual modalities performance, RB fusion consistently exhibits lower word error rates than either of them in any given environment.

The experiments have shown that multimodal fusion consistently outperforms its single-modal counterparts in the majority of environmental conditions, confirming its remarkable robustness. It combines the advantages of both modalities and can be applied in a wider range of environments.

4.5 Experiment 3: Silent Speech

In this portion of the experiment, we conducted a comparison between two fusion strategies and two single-modal approaches in an echoic-only state. Research findings suggest that all fusion strategy can effectively utilize the information provided by the echoic modality, resulting in a significantly lower word error rate compared to using the vocal modality alone. This supports the capacity of our designed dual-modal Keyword Spotting (KWS) system to provide dependable results in environments where user silence is necessary.

4.5.1 Experiment Setup

In this assessment, we employed data from which vocal information had been removed to evaluate the performance of the fusion modality in a silent environment.

Because the experiment relies on echoic information, we continued to use the self-recorded dataset mentioned earlier. We applied the previously mentioned filtering and separation process to all test data, categorizing it into vocal and echoic signals. The vocal segments were substituted with corresponding portions of random silence signals. In this scenario, vocal information is entirely eliminated.

Following that, the echoic and vocal signals are separately fed into their corresponding modalities, resulting in predictions. The fusion modality will generate results based on these predictions and will be compared to both the echoic and vocal modalities to evaluate the system's performance in a silence environment.

4.5.2 Results and Discussion

The experimental results, shown in Fig. 7 (b), display the Word Error Rates (WER) for each modality in the form of bar charts. Various colors on the bars represent the proportions of substitutions, deletions, and insertions. As expected, with the exception of segments in the test set that originally contained silence signals, all signals have been transformed into deletions in the vocal modality, effectively eliminating the possibility of vocal signals providing information. At the same time, the performance of the echoic modality remains unaffected, with error recognition primarily consisting of substitutions and deletions, which aligns with our expectations.

The results indicate that the Word Error Rate (WER) of the two fusion strategies closely approximates that of the echoic modality, which is significantly lower than that of the vocal modality.

In conclusion, the experimental results showcase that our fusion strategy can yield accurate results in an echoic-only environment, a capability that cannot be achieved by the vocal-only single-modal approach. This signifies that our device can operate without requiring additional user intervention, allowing users the freedom to choose whether to speak or use silent speech, and delivers reliable results.

4.6 Experiment 4: Nearby Speaker Interference

In the following analysis, we conducted an investigation into the performance of fusion methods with vocal signals in the presence of interference from other vocal sources. Research findings indicate that the fusion strategy can effectively harness the information from echoic signals. Ultimately, the accuracy of the fusion strategy is within an acceptable range, slightly higher than the performance level of the pure echoic modality and far below that of the vocal modality. This discovery attests that our system can exhibit noteworthy performance even in the face of substantial interference from vocal noise, underscoring its robustness.

4.6.1 Experiment Setup

During this experiment, we simulated the presence of nearby speakers by superimposing the voices of other individuals onto the data, enabling an examination of the system's robustness under these conditions. Only the vocal component was contaminated; the echoic component remained unaffected. We employed the previously described self-recorded dataset for our experimentation, with the vocal data used to introduce interference sourced from the Google Speech Commands v2 dataset [50]. Accounting for differences in average power between the two datasets vocal components, we applied a fixed scaling factor to the vocal signals employed for superimposition to more accurately simulate nearby speakers.

Data from the self-recorded dataset, when superimposed with other human vocal signals, will yield results separately for the silent and vocal single modalities. Predictions from the fusion modality, based on these outcomes, will be compared to assess the system's performance in scenarios involving nearby speakers.



Fig. 7: The performance of our system in three challenging scenarios, measured by Word Error Rate (WER). (a) Fusion consistently achieves the lowest WER, outperforming single modalities in all scenarios. (b) In silent speech, traditional vocal KWS fails entirely, while fusion matches the performance of the echoic modality, notably expanding the system's usage scenarios. (c) In the presence of nearby speakers, the fusion's WER is significantly lower than that of vocal modal systems, greatly reducing false triggers in speech-interference environments.

4.6.2 Results and Discussion

The results of this experiment are illustrated in Fig. 7 (c), where error rates for each modality are presented in the form of bar charts. Different colors of bars represent the proportions of substitution, deletion, and insertion errors. In this context, the vocal modality displayed nearly half of the errors, primarily composed of substitution and deletion errors. In contrast, the echoic modality still produced relatively reliable results, with a certain proportion of substitution errors and very few deletion errors. From the results, it is evident that the fusion modality significantly alleviated the high error rate observed in the vocal modality, including both insertion and substitutions errors.

This experiment demonstrates that even in situations characterized by highly conspicuous background human speech interference, the fusion strategy can effectively prevent false triggers and provide reliable responses. The significant reduction in substitution errors suggests that users can have confidence in the recognition accuracy, even in the presence of human voice noise. Additionally, the near elimination of insertion errors addresses concerns related to unintentional activations in noisy or crowded settings. Users can confidently utilize this approach in environments such as discussions, classrooms, and similar settings without concern for the impact of high-decibel human voices in these scenarios.

4.7 Discussion

Our experimental results demonstrate that our dual-modal approach is more robust compared to both vocal keyword spotting methods and silent speech methods:

Comparison with Vocal KWS method. Our dual-modal approach demonstrates superior performance in various noisy environments compared to conventional keyword spotting methods. Additionally, it can effectively filter out interference from other nearby speakers and supports the use of silent speech when vocalization is inconvenient.

Comparison with Silent Speech method. Our dual-modal approach inherits the capability of prior works on silent speech interfaces. Moreover, in scenarios where silent speech recognition performance experiences significant degradation, such as excessively noisy environments, intense physical activities, and ultrasonic interference, our system can seamlessly leverage vocal speech for recognition. This capability not only aids in avoiding potential system failures but also eliminates the need for manual mode switching.

As a supplement, we have developed a game to validate the experimental results, showcasing the effectiveness of our system in various scenarios. Demonstration video can be viewed at https://youtu.be/fSQoEJ37uEw.

5 LIMITATIONS AND FUTURE WORK

Our work also has some limitations, and we discuss the limitations we have determined through additional studies in this section.

Physical Activities. Walking and head shaking impact the performance of both the vocal and echoic modality pipelines, consequently affecting the overall system. To investigate the influence of these factors, we conducted a series of experiments. The results indicate that: 1) Walking and head shaking have an observable but minor impact on the vocal modality. However, they significantly impact the performance of the echoic modality, causing an increase of around 50% and around 33% in WER, respectively. 2) The fusion method optimized with activities data can mitigate the interference from the echoic modality, resulting in fused WERs that are still lower than the individual WERs of the vocal modality. Nevertheless, physical activities still weaken our system's advantage over conventional vocal KWS. Further enhancement of the echoic modality method remains a potential direction for improvement.

Ultrasonic Interference. Another factor not directly addressed in this paper is the interference from ultrasound emitted by other devices of the same kind. The experiments indicate that: 1) The interference from a stationary ultrasound source is neglectable across multiple directions and distances. This may be because the differential processing in our echoic modality pipeline has a mitigating effect on ultrasound interference. 2) On the other hand, moving ultrasound source has a more significant impact, with an average increase of approximately 7% in WER at a distance of 1m, and closer ultrasound sources causing greater interference. This is because motion diminishes the mitigating effect of the differential processing on interference. Further research into this aspect is a potential avenue for future work.

Other potential future work includes investigating the impact on the performance and user experience of integrating additional modalities in KWS, improving the performance of the echoic modality model in the system, and further minimizing the introduction of additional hardware.

6 CONCLUSION

In this study, we introduce a dual-modal keyword spotting (KWS) system for XR headsets, implemented on the Microsoft HoloLens 2 platform. The key of our system lies in the fusion of features from two distinct modalities: vocal speech and mouth movement information captured through ultrasonic echoes. This integration imparts superior noise robustness and adaptability to diverse scenarios. Specifically,

our approach efficiently utilizes hardware, requiring only off-the-shelf speakers and microphones to obtain information from both modalities. Moreover, our method is computationally lightweight, employing streamlined models and efficient fusion strategies. Our experimental results demonstrate the exceptional performance of this dual-modal system across various challenging scenarios. It outperforms single-modal systems in noisy environments and offers advantages in silent scenarios and situations with nearby speech interference, where traditional vocal KWS systems struggle. Overall, our proposed dual-modal method enhances the noise robustness of KWS systems and notably expands their application scope. This advancement empowers users to engage in speech interactions more frequently, providing not only superior interaction experiences but also enhanced flexibility of choice.

REFERENCES

- A. Bedri, H. Sahni, P. Thukral, T. Starner, D. Byrd, P. Presti, G. Reyes, M. Ghovanloo, and Z. Guo. Toward Silent-Speech Control of Consumer Wearables. *Computer*, 48(10):54–62, Oct. 2015. doi: 10.1109/MC.2015. 310 2
- [2] A. Berg, M. O'Connor, and M. T. Cruz. Keyword Transformer: A Self-Attention Model for Keyword Spotting. In *Interspeech 2021*, pp. 4249– 4253, Aug. 2021. arXiv:2104.00769 [cs, eess]. doi: 10.21437/Interspeech. 2021-1286 4
- [3] L. A. Cheah, J. M. Gilbert, J. A. Gonzalez, P. D. Green, S. R. Ell, R. K. Moore, and E. Holdsworth. A Wearable Silent Speech Interface based on Magnetic Sensors with Motion-Artefact Removal:. In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies*, pp. 56–62. SCITEPRESS - Science and Technology Publications, Funchal, Madeira, Portugal, 2018. doi: 10.5220/ 0006573200560062 2
- [4] G. Chen, C. Parada, and G. Heigold. Small-footprint keyword spotting using deep neural networks. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4087–4091. IEEE, Florence, Italy, May 2014. doi: 10.1109/ICASSP.2014.6854370 2
- [5] T. Chen, B. Steeper, K. Alsheikh, S. Tao, F. Guimbretière, and C. Zhang. C-Face: Continuously Reconstructing Facial Expressions by Deep Learning Contours of the Face with Ear-mounted Miniature Cameras. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pp. 112–125. ACM, Virtual Event USA, Oct. 2020. doi: 10. 1145/3379337.3415879 2
- [6] S. Choi, S. Seo, B. Shin, H. Byun, M. Kersner, B. Kim, D. Kim, and S. Ha. Temporal Convolution for Real-time Keyword Spotting on Mobile Devices, Nov. 2019. arXiv:1904.03814 [cs, eess]. 2, 4
- [7] C. Cioflan, L. Cavigelli, M. Rusci, M. De Prado, and L. Benini. Towards On-device Domain Adaptation for Noise-Robust Keyword Spotting. In 2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS), pp. 82–85. IEEE, Incheon, Korea, Republic of, June 2022. doi: 10.1109/AICAS54282.2022.9869990 2
- [8] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, Aug. 1980. doi: 10.1109/TASSP.1980.1163420 4
- [9] R. Ding, C. Pang, and H. Liu. Audio-Visual Keyword Spotting Based on Multidimensional Convolutional Neural Network. In 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 4138–4142. IEEE, Athens, Oct. 2018. doi: 10.1109/ICIP.2018.8451096 1
- [10] M. Elepfandt and M. Grund. Move it there, or not?: the design of voice commands for gaze with speech. In *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, pp. 1–3. ACM, Santa Monica California, Oct. 2012. doi: 10.1145/2401836.2401848 2
- [11] I. Fung and B. Mak. End-To-End Low-Resource Lip-Reading with Maxout Cnn and Lstm. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2511–2515. IEEE, Calgary, AB, Apr. 2018. doi: 10.1109/ICASSP.2018.8462280 1
- [12] M. J. F. Gales, K. M. Knill, and A. Ragni. Low-Resource Speech Recognition and Keyword-Spotting. In A. Karpov, R. Potapova, and I. Mporas, eds., *Speech and Computer*, vol. 10458, pp. 3–19. Springer International Publishing, Cham, 2017. Series Title: Lecture Notes in Computer Science. doi: 10.1007/978-3-319-66429-3_1 2
- [13] Y. Gao, Y. Jin, J. Li, S. Choi, and Z. Jin. EchoWhisper: Exploring an Acoustic-based Silent Speech Interface for Smartphone Users. Pro-

ceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(3):1–27, Sept. 2020. doi: 10.1145/3411830 1, 2

- [14] A. Grinshpoon, S. Sadri, G. J. Loeb, C. Elvezio, and S. K. Feiner. Hands-Free Interaction for Augmented Reality in Vascular Interventions. In 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 751–752. IEEE, Reutlingen, Mar. 2018. doi: 10.1109/VR.2018.8446259 2
- [15] R. Hofe, S. R. Ell, M. J. Fagan, J. M. Gilbert, P. D. Green, R. K. Moore, and S. I. Rybchenko. Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. *Speech Communication*, 55(1):22–32, Jan. 2013. doi: 10.1016/j.specom.2012.02.001 2
- [16] J. Hombeck, H. Voigt, T. Heggemann, R. R. Datta, and K. Lawonn. Tell Me Where To Go: Voice-Controlled Hands-Free Locomotion for Virtual Reality Systems. In 2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR), pp. 123–134. IEEE, Shanghai, China, Mar. 2023. doi: 10. 1109/VR55154.2023.00028 2
- [17] A. Kapur, S. Kapur, and P. Maes. AlterEgo: A Personalized Wearable Silent Speech Interface. In 23rd International Conference on Intelligent User Interfaces, IUI '18, pp. 43–53. Association for Computing Machinery, New York, NY, USA, Mar. 2018. doi: 10.1145/3172944.3172977 1, 2
- [18] M. Kaur, M. Tremaine, N. Huang, J. Wilder, F. Flippo, and S. Mantravadi. Where is "it"? Event Synchronization in Gaze-Speech Input Systems. 2
- [19] B. Kim, S. Chang, J. Lee, and D. Sung. Broadcasted Residual Learning for Efficient Keyword Spotting, July 2023. arXiv:2106.04140 [cs, eess]. 2, 4
- [20] N. Kimura, T. Gemicioglu, J. Womack, R. Li, Y. Zhao, A. Bedri, A. Olwal, J. Rekimoto, and T. Starner. Mobile, Hands-free, Silent Speech Texting Using SilentSpeller. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–5. ACM, Yokohama Japan, May 2021. doi: 10.1145/3411763.3451552 2
- [21] N. Kimura, K. Hayashi, and J. Rekimoto. TieLent: A Casual Neck-Mounted Mouth Capturing Device for Silent Speech Interaction. In Proceedings of the International Conference on Advanced Visual Interfaces, pp. 1–8. ACM, Salerno Italy, Sept. 2020. doi: 10.1145/3399715.3399852 2
- [22] N. Kimura, M. Kono, and J. Rekimoto. SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–11. ACM, Glasgow Scotland Uk, May 2019. doi: 10.1145/3290605.3300376 1, 2
- [23] R. Kumar, V. Yeruva, and S. Ganapathy. On Convolutional LSTM Modeling for Joint Wake-Word Detection and Text Dependent Speaker Verification. In *Interspeech 2018*, pp. 1121–1125. ISCA, Sept. 2018. doi: 10. 21437/Interspeech.2018-1759 2
- [24] Y. Kunimi, M. Ogata, H. Hiraki, M. Itagaki, S. Kanazawa, and M. Mochimaru. E-MASK: A Mask-Shaped Interface for Silent Speech Interaction with Flexible Strain Sensors. In *Augmented Humans* 2022, pp. 26–34. ACM, Kashiwa, Chiba Japan, Mar. 2022. doi: 10.1145/3519391.3519399 1, 2
- [25] K.-S. Lee. EMG-Based Speech Recognition Using Hidden Markov Models With Global Control Variables. *IEEE Transactions on Biomedical Engineering*, 55(3):930–940, Mar. 2008. doi: 10.1109/TBME.2008.915658 1, 2
- [26] M. Lee, M. Billinghurst, W. Baek, R. Green, and W. Woo. A usability study of multimodal input in an augmented reality environment. *Virtual Reality*, 17(4):293–305, Nov. 2013. doi: 10.1007/s10055-013-0230-0 2
- [27] K. Li, R. Zhang, B. Liang, F. Guimbretière, and C. Zhang. EarIO: A Low-power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–24, July 2022. doi: 10. 1145/3534621 4
- [28] R. Li, J. Wu, and T. Starner. TongueBoard: An Oral Interface for Subtle Input. 2019. 2
- [29] I. Lopez-Espejo, Z.-H. Tan, J. H. L. Hansen, and J. Jensen. Deep Spoken Keyword Spotting: An Overview. *IEEE Access*, 10:4169–4199, 2022. doi: 10.1109/ACCESS.2021.3139508 2, 4
- [30] I. Lopez-Espejo, Z.-H. Tan, and J. Jensen. A Novel Loss Function and Training Strategy for Noise-Robust Keyword Spotting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2254–2266, 2021. doi: 10.1109/TASLP.2021.3092567 2
- [31] A. H. Michaely, X. Zhang, G. Simko, C. Parada, and P. Aleksic. Keyword spotting for Google assistant using contextual speech recognition. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop

(ASRU), pp. 272–278. IEEE, Okinawa, Dec. 2017. doi: 10.1109/ASRU. 2017.8268946 2

- [32] P. Monteiro, G. Goncalves, H. Coelho, M. Melo, and M. Bessa. Handsfree interaction in immersive virtual reality: A systematic review. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2702–2713, May 2021. doi: 10.1109/TVCG.2021.3067687 2
- [33] T. Piumsomboon, D. Altimira, H. Kim, A. Clark, G. Lee, and M. Billinghurst. Grasp-Shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality. In 2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 73–82. IEEE, Munich, Germany, Sept. 2014. doi: 10.1109/ISMAR. 2014.6948411 2
- [34] G. Pomianos, C. Neti, G. Gravier, A. Garg, and A. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, Sept. 2003. doi: 10.1109/JPROC.2003.817150 1, 4
- [35] R. Prabhavalkar, R. Alvarez, C. Parada, P. Nakkiran, and T. N. Sainath. Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4704–4708. IEEE, South Brisbane, Queensland, Australia, Apr. 2015. doi: 10.1109/ICASSP.2015.7178863 2
- [36] H.-R. Rantamaa, J. Kangas, M. Jordan, H. Mehtonen, J. Mäkelä, K. Ronkainen, M. Turunen, O. Sundqvist, I. Syrjä, J. Järnstedt, and R. Raisamo. Evaluation of voice commands for mode change in virtual reality implant planning procedure. *International Journal of Computer Assisted Radiology and Surgery*, 17(11):1981–1989, June 2022. doi: 10. 1007/s11548-022-02685-1 2
- [37] J. Rekimoto and Y. Nishimura. Derma: Silent Speech Interaction Using Transcutaneous Motion Sensing. In *Augmented Humans Conference 2021*, pp. 91–100. ACM, Rovaniemi Finland, Feb. 2021. doi: 10.1145/3458709. 3458941 1, 2
- [38] J. Rohlicek, W. Russell, S. Roukos, and H. Gish. Continuous hidden Markov modeling for speaker-independent word spotting. In *International Conference on Acoustics, Speech, and Signal Processing*, pp. 627–630. IEEE, Glasgow, UK, 1989. doi: 10.1109/ICASSP.1989.266505 2
- [39] R. Rose and D. Paul. A hidden Markov model based keyword recognition system. In *International Conference on Acoustics, Speech, and Signal Processing*, pp. 129–132. IEEE, Albuquerque, NM, USA, 1990. doi: 10. 1109/ICASSP.1990.115555 2
- [40] A. Rosenberg, K. Audhkhasi, A. Sethy, B. Ramabhadran, and M. Picheny. End-to-end speech recognition and keyword search on low-resource languages. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5280–5284. IEEE, New Orleans, LA, Mar. 2017. doi: 10.1109/ICASSP.2017.7953164 2
- [41] T. N. Sainath and C. Parada. Convolutional neural networks for smallfootprint keyword spotting. In *Interspeech 2015*, pp. 1478–1482. ISCA, Sept. 2015. doi: 10.21437/Interspeech.2015-352 2
- [42] B. Shi, W.-N. Hsu, and A. Mohamed. Robust Self-Supervised Audio-Visual Speech Recognition, July 2022. arXiv:2201.01763 [cs, eess]. 1
- [43] M. Sun, A. Raju, G. Tucker, S. Panchapagesan, G. Fu, A. Mandal, S. Matsoukas, N. Strom, and S. Vitaladevuni. Max-pooling loss training of long short-term memory networks for small-footprint keyword spotting. In 2016 IEEE Spoken Language Technology Workshop (SLT), pp. 474–480. IEEE, San Diego, CA, Dec. 2016. doi: 10.1109/SLT.2016.7846306 2
- [44] R. Tang and J. Lin. Deep Residual Learning for Small-Footprint Keyword Spotting, Sept. 2018. arXiv:1710.10361 [cs]. 2, 4
- [45] J. Thiemann, N. Ito, and E. Vincent. The Diverse Environments Multichannel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings. pp. 035081–035081. Montreal, Canada, 2013. doi: 10.1121/1.4799597 5
- [46] R. Vygon and N. Mikhaylovskiy. Learning Efficient Representations for Keyword Spotting with Triplet Loss. In A. Karpov and R. Potapova, eds., *Speech and Computer*, vol. 12997, pp. 773–785. Springer International Publishing, Cham, 2021. Series Title: Lecture Notes in Computer Science. doi: 10.1007/978-3-030-87802-3_69 4
- [47] T. Wang, D. Zhang, Y. Zheng, T. Gu, X. Zhou, and B. Dorizzi. C-FMCW Based Contactless Respiration Detection Using Acoustic Signal. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1(4):1–20, Jan. 2018. doi: 10.1145/3161188 4
- [48] Y. Wang, M. Zhang, R. Wu, H. Gao, M. Yang, Z. Luo, and G. Li. Silent Speech Decoding Using Spectrogram Features Based on Neuromuscular Activities. *Brain Sciences*, 10(7):442, July 2020. doi: 10.

3390/brainsci10070442 1, 2

- [49] Z. Wang, H. Wang, H. Yu, and F. Lu. Interaction With Gaze, Gesture, and Speech in a Flexibly Configurable Augmented Reality System. *IEEE Transactions on Human-Machine Systems*, 51(5):524–534, Oct. 2021. doi: 10.1109/THMS.2021.3097973 2
- [50] P. Warden. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition, Apr. 2018. arXiv:1804.03209 [cs]. 5, 6, 7
- [51] J. Wilpon, L. Miller, and P. Modi. Improvements and applications for key word recognition using hidden Markov modeling techniques. In [Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing, pp. 309–312 vol.1. IEEE, Toronto, Ont., Canada, 1991. doi: 10.1109/ICASSP.1991.150338 2
- [52] K. Xu, D. Li, N. Cassimatis, and X. Wang. LCANet: End-to-End Lipreading with Cascaded Attention-CTC. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 548–555. IEEE, Xi'an, May 2018. doi: 10.1109/FG.2018.00088 1
- [53] M. Xu and X.-L. Zhang. Depthwise Separable Convolutional ResNet with Squeeze-and-Excitation Blocks for Small-Footprint Keyword Spotting. In *Interspeech 2020*, pp. 2547–2551. ISCA, Oct. 2020. doi: 10.21437/ Interspeech.2020-1045 2
- [54] R. Zhang, M. Chen, B. Steeper, Y. Li, Z. Yan, Y. Chen, S. Tao, T. Chen, H. Lim, and C. Zhang. SpeeChin: A Smart Necklace for Silent Speech Recognition. *Proceedings of the ACM on Interactive, Mobile, Wear-able and Ubiquitous Technologies*, 5(4):1–23, Dec. 2021. doi: 10.1145/ 3494987 2
- [55] R. Zhang, K. Li, Y. Hao, Y. Wang, Z. Lai, F. Guimbretière, and C. Zhang. EchoSpeech: Continuous Silent Speech Recognition on Minimallyobtrusive Eyewear Powered by Acoustic Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1– 18. ACM, Hamburg Germany, Apr. 2023. doi: 10.1145/3544548.3580801 1, 2, 3, 4
- [56] Y. Zhang, Y.-C. Chen, H. Wang, and X. Jin. CELIP: Ultrasonic-based Lip Reading with Channel Estimation Approach for Virtual Reality Systems. In Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers, pp. 580–585. ACM, Virtual USA, Sept. 2021. doi: 10.1145/3460418.3480163 1, 2
- [57] P. Zhou, W. Yang, W. Chen, Y. Wang, and J. Jia. Modality Attention for End-to-end Audio-visual Speech Recognition. In *ICASSP 2019 -2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6565–6569. IEEE, Brighton, United Kingdom, May 2019. doi: 10.1109/ICASSP.2019.8683733 1
- [58] Y. Zhuang, X. Chang, Y. Qian, and K. Yu. Unrestricted Vocabulary Keyword Spotting Using LSTM-CTC. In *Interspeech 2016*, pp. 938–942. ISCA, Sept. 2016. doi: 10.21437/Interspeech.2016-753 2