Unvoiced: Designing an LLM-assisted Unvoiced User Interface using Earables

Tanmay Srivastava
tsrivastava@cs.stonybrook.edu
Stony Brook University
New York, USA

Prerna Khanna
pkhanna@cs.stonybrook.edu
Stony Brook University
New York, USA

Shijia Pan
span24@ucmerced.edu
University of California, Merced
Merced, USA

Phuc Nguyen
vp.nguyen@cs.umass.edu
University of Massachusetts Amherst
Amherst, USA

Shubham Jain
jain@cs.stonybrook.edu
Stony Brook University
New York, USA

Abstract

We present *Unvoiced*, a novel unvoiced user interface that leverages jaw motion to enable users to silently interact with their devices using earables. The core idea is to translate low-frequency jaw motion signals into high-frequency information-rich mel spectrograms. Our proposed cross-modal translation incorporates phonetic, contextual, and syntactic information, while the specialized loss function optimizes for these linguistic features. This ensures that the generated spectrograms capture nuanced speech characteristics. Evaluated for 19 users across four tasks, *Unvoiced* demonstrates >94% task completion rate and <9% word error rate for over 90% of phrases. Further, *Unvoiced* maintains >90% task completion rate in noisy conditions.

CCS Concepts

• **Human-centered computing** → **Accessibility technologies**; **Interaction techniques**.

Keywords

Accessible Interfaces, Silent Speech, Transformers, Earables, IMU Sensing, GPT, LLM

ACM Reference Format:
Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2024. *Unvoiced*: Designing an LLM-assisted Unvoiced User Interface using Earables. In *ACM Conference on Embedded Networked Sensor Systems (SenSys '24), November 4–7, 2024, Hangzhou, China.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3666025.3699374

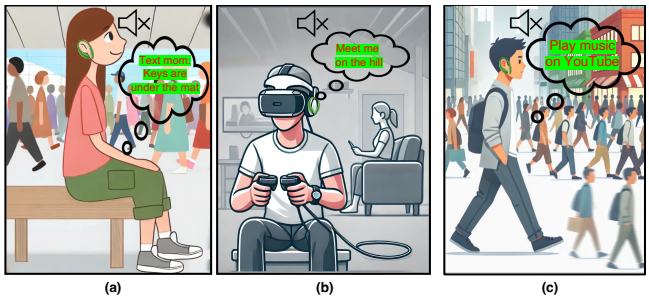


Figure 1: Use cases for *Unvoiced*. (a) Discreet interaction in public spaces. (b) VR interaction for a better immersive experience. (c) Hands-free and reliable voice interaction in a noisy background.

1 Introduction

Speech-based interactions are everywhere. In fact, Voice User Interfaces (VUIs) have become the de facto interaction modality for intelligent assistants (e.g. Alexa [3] or Siri [6]), Virtual Reality (VR) [58, 59], in-vehicle interactions [41, 91], smart home devices [4, 101], and other Internet of Things (IoT) devices [16, 68] in both, public and private spaces. Despite its success and popularity as an interaction modality, audible speech has severe limitations in practical environments. VUIs are often not robust in hearing or understanding the user in the presence of background noise, can compromise privacy in public spaces (consider dictating a private text in a public space), or can simply be impractical in quiet environments where it is desirable to not disturb cohabitants (such as in the library, classroom, or even at home).

To address these limitations, unvoiced user interfaces (UUI), also known as silent speech interfaces (SSIs), have recently garnered significant interest [64, 80]. Unlike VUIs, UUIs enable users to communicate without vocalization, broadening their applicability to situations where conventional speech is impractical. A contemporary study [65] revealed that UUIs are generally perceived as more socially acceptable than voiced speech, with users showing greater tolerance for errors. Research has also highlighted that social discomfort and privacy issues significantly influence users' perceptions and willingness to adopt voice assistants [94]. By eliminating the need for audible speech, UUIs offer enhanced privacy protection. These benefits position them as a promising technology for expanding the functionality of voice assistants through silent interactions.

While interesting, most existing work in this space is difficult to realize for widespread adoption. Works that rely on camera-based

SenSys '24, November 4–7, 2024, Hangzhou, China

ACM ISBN 979-8-4007-0697-4/24/11
https://doi.org/10.1145/3666025.3699374

lip reading [26, 62, 63, 67, 86, 102] require unobstructed camera view, hampering user's privacy and limiting system usability. Other systems that employ acoustic sensing face challenges due to non-hands-free operation, privacy concerns, and often require the user to be stationary or hold a mobile device [22, 33, 89]. In the realm of wearable devices, many solutions to recognizing silent speech involve placing sensors inside the oral cavity or on/around the lips [43, 96], which may be both, physically uncomfortable as well as socially unacceptable. Recent research efforts have focused on positioning less intrusive sensors in less visible locations, such as behind the ear or under the chin [39, 44, 83, 107]. However, they often necessitate additional user efforts like slowed speech to maintain performance or can only recognize isolated phonemes or words. *It is evident that there is a need for an unvoiced interface that allows users to seamlessly communicate with their devices privately, without the impact of background noise, but also without using obtrusive sensing devices that are not socially acceptable.*

We design *Unvoiced*, an unvoiced user interface that can enable silent interactions ubiquitously. We achieve a robust unvoiced interface that can work across multiple applications by leveraging the rising trend of sensor-equipped earables [78]. *Unvoiced* is an IMU-based system that tracks the user's jaw motion as they articulate without vocalizing to recognize silent speech. Recent work on using jaw motion to recognize silent speech has demonstrated immense promise [83, 106]. We leverage the unobtrusive around-the-ear wearable prototype [39] to facilitate *Unvoiced*. However, since the jaw is a secondary articulator, deciphering speech from jaw motion alone is challenging. With the recent success of Large Language Models (LLMs), the question we then ask is: *Is it possible to harness the capabilities of LLMs to support unvoiced user interactions using jaw motion only?* To design this LLM-assisted robust UUI, our key intuition is to translate IMU data to audio spectrograms which can then be converted to text using off-the-shelf LLMs. Figure 1 demonstrates several use-case scenarios.

Challenges. Developing *Unvoiced* entails 3 key challenges:

(1) *Cross-modal translation faces significant data domain variance.* Translating from jaw motion to Mel spectrograms is challenging because the source modality (jaw motion) is sampled at a much lower frequency than the target modality (audio spectrograms), which is sampled at $10^3\times$ IMU data. Moreover, spectrograms encapsulate rich features from audible speech (such as formants and chroma features), which are challenging to learn from jaw motion only.

(2) *Jaw motion and speech are not directly related.* Speech results from a high degree of overlap and continuous movements of multiple articulators [53]. The jaw is a secondary articulator with few degrees of freedom and is not actively involved in speech production (multiple sounds result in the same jaw movement). As a result, generating rich audio spectrograms from the limited measurements of jaw motion is challenging.

(3) *Variable context across target applications.* Applications that would benefit from UUI have very different contexts and there exist multiple possibilities for what is considered a valid interaction phrase within the scope of each application. For example, a voice assistant can accept both "Add apples to shopping list" and "Delete apples from shopping list" as a valid command. However, errors in detecting even one word can completely alter the task output.

Contributions. *Unvoiced* addresses these challenges by first modeling the relationship between jaw motion and Mel spectrograms using a multi-modal Transformer model with custom losses that incorporate phonetic knowledge into the model. This spectrogram can now be used with any LLM to generate text. However, LLMs are large and can produce valid outputs outside the context of the current application. We scope their output by injecting contextual information, derived from phonetic components. Finally, we demonstrate *Unvoiced* on 4 applications with different contexts and accuracy needs. Results reveal that our generated spectrograms are very close to the target spectrograms. We can achieve task completion accuracy >94%. We realize *Unvoiced* using a twin-IMU earable prototype that was recently developed [83] and has been used in other works successfully [84]. This prototype allows us to extract jaw motion while cancelling out motion-induced noise.

To summarize, we make the following contributions:

- To the best of our knowledge, this is among the earliest attempts to perform cross-modal translation to generate Mel spectrograms (high frequency, rich information) from low-frequency ear-mounted IMU sensors.
- This is the first unvoiced user interface that brings the advances in LLMs to develop a robust unvoiced interaction modality relying only on a secondary articulator.
- We refine the large space of outputs from LLMs by leveraging phonetic components leading to contextually correct interaction phrases.
- Extensive evaluations for 19 users across 4 applications and 3 baseline techniques, achieving a task completion rate of 94.3%.

Applications. *Unvoiced* aims to enable seamless and private interactions with various voice-enabled applications. It allows users to silently control media playback and smart home devices, enhancing privacy and security by preventing eavesdropping and impersonation attacks. The system can facilitate secure transactions using silent speech as a second-factor authentication [84] and enables discreet note-taking and messaging in professional and educational settings. In VR/AR environments, it provides hands-free control, enhancing immersive experiences without disrupting the real world. Importantly, it has the potential to offer an inclusive alternative for individuals with speech impairments, enabling technology interaction in a hands-free voice-free manner.

2 LLM-Assisted User Interfaces

2.1 Background

Jaw as a cue for speech production. We explore the possibility of developing a continuous unvoiced speech recognition system that utilizes jaw motion as the primary input. Despite being a secondary articulator compared to the lips, teeth, tongue, and palates involved in sound production, the jaw facilitates its movements without directly affecting airflow for sound distinction. Specifically, the temporomandibular joint (TMJ), which allows rotational and translational movements of the jawbone, supports the necessary movements of the lips and tongue for articulating words. Therefore, by tracking the *jaw—an articulator associated with speech but not directly involved in sound production—* we enable continuous unvoiced speech recognition. This possibility has been confirmed by prior works in phoneme detection [39, 83, 106] for isolated sounds

and commands, skin's displacement monitoring [108] for numeric passcodes, and cheek movement monitoring [45, 46] to enable facial gesture recognition.

Leveraging LLMs in User Interface Design: VUI techniques have been significantly improved by advancements in Language Model (LLM) capabilities, particularly to bridge the gaps between speech recognition and text generation [42]. Recent LLMs—those based on Transformer architectures—have vastly improved the accuracy of converting human speech into text [60]. By leveraging large-scale data and sophisticated ML techniques, LLMs enhance VUI systems' ability to understand diverse contexts even with accents in multiple languages [40]. LLM-powered VUI not only enriches user interactions by making them more intuitive and responsive but also expands the application of voice-driven technology across industries, from virtual assistants to automated customer service and many others [12, 25, 54, 87]. Following the success of LLMs in enhancing VUIs, we aim to leverage LLMs in our unvoiced user interface design.

2.2 Preliminary Study

To validate the feasibility of LLM-assisted UUI, we conducted a preliminary study where a user wore the twin-IMU prototype and spoke the sentence "Set alarm for six A.M.", at slowed and normal speech rates of 30 and 55 WPM (words per minute) respectively (see Figure 2). We observe that at 30 WPM, [83] was able to recognize phonetic components and perform recognition. However, at normal speech rate, multiple important phonetic features are suppressed in the jaw motion signal. This leads to sparse information about phonetic components and few words are recognized compared to [83]. We feed these recognized words directly into an LLM and observe the outcomes. Since only 2 words ("alarm" and "six") could be recognized, the LLM could not provide the desired response, i.e. the complete phrase. This attempt to develop an LLM-assisted UUI using jaw motion was unsuccessful due to limited speech information. From this observation, the key research question of this work becomes "*How can robust spectrograms, which are essential inputs for LLM-powered speech interfaces, be created from jaw motion data?*". We aim to generate spectrograms as they are widely used in state-of-the-art open-source speech recognition systems such as Whisper [72], DeepSpeech2 [5], and DeepSpeech [71], However, reliable speech spectrograms are challenging to obtain from IMU due to low frequency and relatively less speech information in jaw motion.

3 System Overview

Unvoiced enables silent interactions by combining jaw motion tracking and deep learning model with a custom loss function. IMU captures the subtle movements of the jaw during the silent speech, providing a low-frequency data stream that is converted to high-frequency mel-spectrograms. Figure 3 presents an overview of *Unvoiced*.

During the training phase, synchronized jaw motion, audio data, and text are used to train a deep-learning model to generate spectrograms. Once trained, in the inference phase the model relies on jaw motion data to generate the audio spectrogram and, convert it to text.

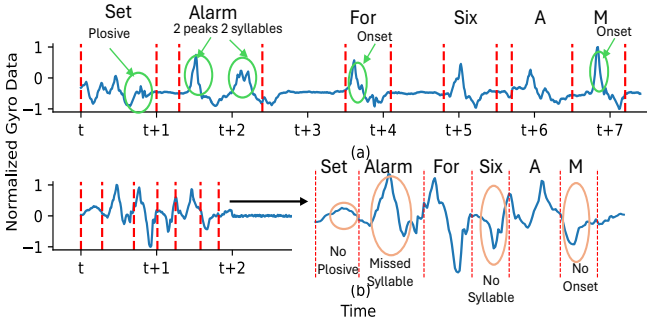


Figure 2: Example jaw motions when a user speaks the phrase "Set alarm for six AM" at 30 WPM and 55 WPM.

This translation is inherently complex due to the sparse nature of the jaw motion data and the high-fidelity spectrogram required for speech recognition. To simplify the task for the model, we propose a phrase segmentation module (§ 4), which breaks down continuous phrase data into words. The segmentation significantly reduces the sequence length of the outputs, thereby not only simplifying the translation task but also increasing the granularity and number of training samples available.

Despite the task simplification through phrase segmentation, the goal of interpreting subtle jaw movements to reconstruct spectral features remains challenging. To solve this challenge, we propose a novel transformer-based encoder-decoder model equipped with custom-designed loss functions. After the phrases have been segmented into words, the inputs for the training model are prepared which are the muti-view for the 6-axis IMU word data, the weighted frequency mask of the audio spectrogram, and the GPT2 text embeddings for the word (§ 5.1). The encoder-decoder model is trained with these inputs using custom loss functions (§ 5.3). These loss functions describe the linguistic properties, such as phonetic, prosodic, and syntactic information from the limited articulatory data, while minimizing spectral divergence. This novel approach ensures that the generated spectrogram follows the implicit linguistic rules and achieves a close representation of the intended speech output (§ 5.3).

During the inference phase, jaw motion data associated with silent speech is captured, the same pre-processing as in the training phase is applied, and it is segmented into words. Our trained model then generates a spectrogram for each word. The final step involves converting these spectrograms into text with the Whisper speech recognition system. In instances of missing information like skipped words due to non-recognizable jaw motion, we employ generative pre-trained transformers (GPT) to produce contextually accurate phrases, further refined by incorporating phonetic information (§ 6). This integration not only showcases the practical utility of our system but also highlights its compatibility with existing speech-recognition methods.

4 Phrase Segmentation

The goal of phrase segmentation algorithm is to segment the IMU signal for a phrase into words after removing noise. Segmenting the phrase into its constituent words helps us twofold: (1) simplifies the task of spectrogram generation by allowing the model to generate

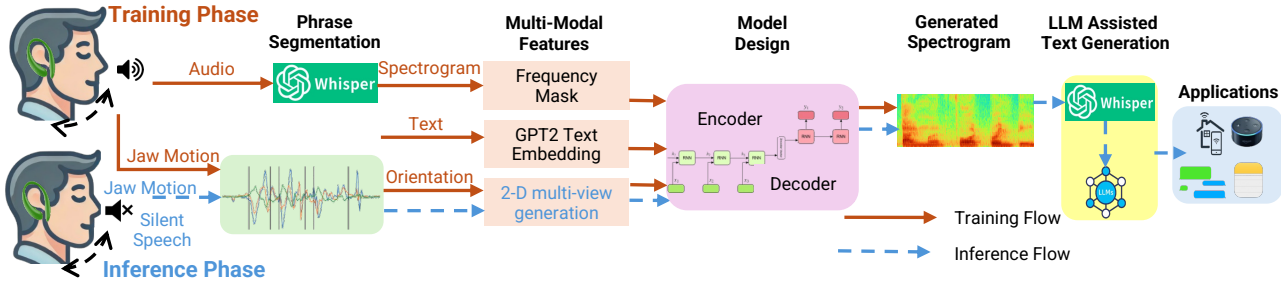


Figure 3: *Unvoiced* Overview

shorter spectrogram signals compared to entire phrases; (2) allow us to train the model for individual words. Word-level processing grants direct access to lexical information like the location of a linguistic unit (word) that helps in semantic understanding. Recall that *Unvoiced* utilizes a twin-IMU noise cancellation design [83, 84], with IMUs placed at the temporal bone and temporomandibular joint (TMJ). We employ an adaptive FIR filter to remove head and body movement artifacts from the TMJ signal. We remove the effect of gravity and DC bias from accelerometer data, followed by a 25 Hz cutoff third-order low-pass filter to extract jaw motion [84].

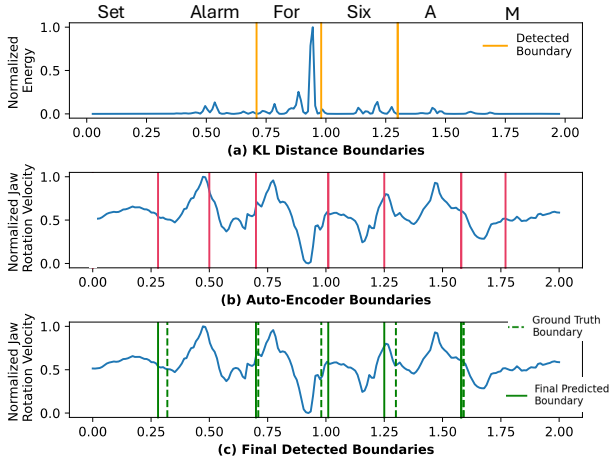


Figure 4: Phrase segmentation. (a) KL distance-based method, which underestimates word counts. (b) Autoencoder method, which overestimates word counts. (c) Final detected word boundaries (solid lines) compared with ground truth (dotted lines).

Word boundaries often coincide with the start and end of syllables and reveal pauses, during which the jaw returns to its starting position. While identifying phonemic components, such as syllables, plosives, etc. can be easily performed at lower speech rates and for isolated words (Figure 2 [83]), at normal speech rates, the jaw does not always come to its starting or resting position between words. As a result, many of these features indicating word boundaries are suppressed in our data.

To address this, we propose a hybrid method for word boundary detection that combines signal processing (Kullback-Leibler-Based (KL) [35, 70] with machine learning solutions (autoencoder anomaly detection). This approach leverages the KL method's ability to identify significant changes in jaw motion and the autoencoder's sensitivity to subtle variations. By integrating these techniques, we address the underestimation of word boundaries by KL distance alone (Figure 4a) and the overestimation by the autoencoder (Figure 4b), resulting in more accurate detection across diverse speech patterns and jaw motions (Figure 4c).

Figure 4(a) illustrates the word boundaries identified by the KL distance for the phrase "Set alarm for six AM". However, this approach fails for smaller words or those with minimal jaw motion, such as "set" and "M". We notice that since the degree of jaw motion can vary with the content for each user, KL distance-based segmentation underestimates the number of words. To address this shortcoming, we develop a lightweight autoencoder architecture to detect word boundaries[31, 47, 52], shown in Figure 4(b). We train the model on 50 ms normalized gyroscope z-axis windows using mean squared error as loss. We extract the word boundaries by inputting segmented windows into the model and marking windows with high reconstruction loss. As we observe, the autoencoder detects numerous false positives, possibly because there can be a change in the jaw motion within a word the model has not seen in training and flags it as an anomaly (window).

To overcome the limitations of both approaches, we propose a hybrid method that leverages their complementary nature. To remove false peaks from the autoencoder output, we identify peaks within 0.2-seconds of each other, and then look for corresponding detections from KL to find a match. Peaks that are further apart, are accepted as detected boundaries. Figure 4(c) shows the final detected word boundaries along with ground truth boundaries, manually extracted from slowed-down audio.

5 Cross-modal Translation: Jaw motion to Spectrogram

In this section, we detail our design to transform jaw motion into spectrograms. Our decision to reconstruct spectrograms rather than directly translating jaw motion to text is driven by several critical factors. First, the dimensional disparity between low-frequency jaw motion data and high-dimensional text embeddings presents a significant challenge for direct translation and requires substantial training data [21]. Spectrograms serve as an ideal intermediate representation, bridging this gap by encoding rich temporal and

frequency information, an approach explored in related unvoiced interface works [19, 20]. Furthermore, spectrograms preserve crucial speech characteristics such as formant structures, pitch contours, and temporal dynamics, which are essential for accurate speech recognition but difficult to capture directly from jaw motion [24]. This approach also allows us to leverage state-of-the-art speech recognition models like Whisper, which are optimized for spectrogram inputs, without extensive retraining. Lastly, this modular architecture enhances flexibility, allowing for independent improvements in either the jaw-to-spectrogram or spectrogram-to-text components as new techniques emerge. It also facilitates noise robustness [104] and cross-lingual adaptability [7], while opening possibilities for data augmentation techniques like SpecAugment [66]. Building upon these advantages of IMU to spectrogram translation, we now delve into the specifics of our jaw motion to spectrogram translation model, including the multi-modal input streams and the architecture, followed by our custom loss functions.

5.1 Input Feature Extraction

The inputs to the model include 6-axis jaw motion data, audio time-frequency domain features, and textual information.

Jaw motion features. From the segmented words, we only retain those words that have significant jaw motion by determining if at least 20% of the values are greater than the mean value of the window [61]. Next, we extract the spatial and geometric information from the raw 6-axis IMU data by computing jaw motion orientation using the Kalman filter and then convert them to isometric views [103], which capture the jaw movement in the 2D plane from different viewpoints. They allow for maintaining the true scale of the motion throughout the view and have been used in previous works [11, 90]. We define 6 rotations to generate the isometric views: primary rotation is defined with an angle of 45 degrees and five additional rotation matrices are defined using combinations of angles (-45, 135, -135, 180, and -180) to capture diverse perspectives. These isometric views are provided as inputs to the model.

Audio features We incorporate two audio features into our model: (1) spectral information and (2) phonetic information. For spectral features we use the spectrogram of the voiced speech. We extract word boundaries for the audio data, using OpenAI's Whisper [72]. After segmenting the phrase spectrogram into words we low pass filter the spectrogram to retain usable voice frequency ranges(300 to 3400 Hz) [77]. This reduces the size of the output spectrogram simplifying the task of spectrogram generation. Since jaw motion has been shown to capture phonetic information, we learn similar information from audio data, such as the number of syllables and the presence of plosives. For this, we utilize the envelope of the audio time-domain signal which exhibits rising and falling peaks of energy that correspond to the number of syllables in the speech [17, 49]. Plosives manifest as sudden bursts of energy in the time-domain signal, influencing the overall magnitude [36]. By incorporating the audio time-domain envelope, we hypothesize that the model can learn to correlate these phonetic features with IMU signal, thereby improving its ability to generate accurate spectrograms.

Textual and Syntactic Features. Finally, we leverage the GPT-2 language model [74] to extract word embeddings for the corresponding text of each audio file. These embeddings provide contextual information that enables the model to identify instances of the same word across different training samples, hence providing consistent performance across sessions. Lastly, we incorporate syntactic information by adding a location token to the model, indicating the position of each word within the sentence (i.e., beginning, middle, or end). The model can leverage this location token to predict intonation patterns, which vary depending on the word's position in the sentence.

5.2 Model Design

Unvoiced's model architecture consists of an encoder, a decoder with attention, and a GPT-2 embedding layer.

Encoder. The encoder takes the jaw motion features as input, represented as a sequence of vectors. We pad the IMU and the audio data with zeros to be of equal length of 6 seconds. Then the IMU data is passed through a multi-layer bidirectional LSTM (BiLSTM) network to capture the temporal dependencies and extract features. The BiLSTM layers in the encoder have 512 hidden units each, and multiple layers are stacked to capture more complex temporal patterns. Additionally, dense layers with ReLU activation are inserted between the LSTM layers to learn abstract representations. The final hidden states of the forward and backward LSTMs are concatenated using a dense layer to obtain the encoder's output representations

Decoder with Multi-Head Attention. The decoder is a multi-layer BiLSTM network with 512 hidden units in each layer. We use multi-head attention in the decoder to capture the complex relationships between jaw motion and speech features. Additionally, it helps the decoder attend to relevant parts of the input sequence at different time steps, which is crucial for generating accurate spectrograms from jaw motion. Similar to the encoder, we stack multiple LSTM layers, and dense layers with ReLU activation. The decoder takes the target audio spectrogram as input during training and predicts them from jaw motion during inference. The multi-head attention mechanism allows the model to capture more complex and diverse relationships between the input and output sequences, enabling the generation of high-quality Mel features.

GPT-2 Text Embedding Layer. Textual features allow the model to tag the jaw motion-spectrogram input pairs with labels helping the model to generate generalizable spectrograms. For instance, the spectrograms for the word "Alarm" from different sessions should be close to each other. We use GPT2 text embeddings to incorporate textual information into the model as they are trained on similar datasets as with open-to-use prompt-based LLMs (GPT-4). For each phase, the corresponding text data is tokenized and padded to a fixed length of 512. The tokenized text is then passed through a pre-trained GPT-2 model to obtain text embeddings, which are then repeated to match the sequence length of the decoder outputs using a RepeatVector layer. The repeated text embeddings are concatenated with the multi-head attention output along the feature dimension to form a combined context vector. This allows the model to condition the audio feature generation process on both the IMU data and the textual input.

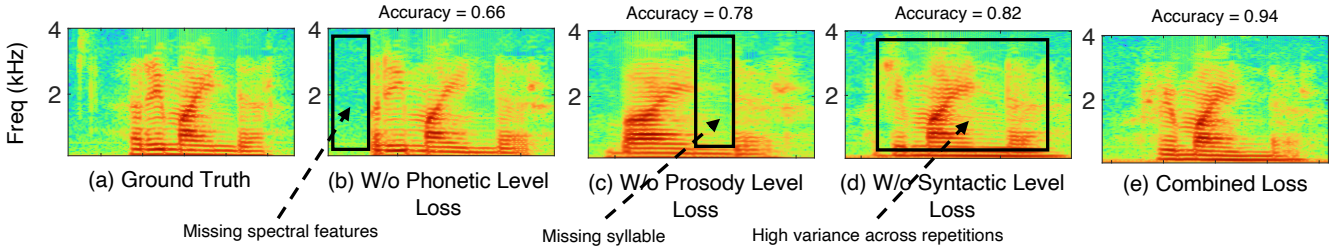


Figure 5: Information captured by each loss function for the word "Reminder".

Output Layer and Training. The combined context vector, which contains information from the IMU data, target audio features, and textual input, is passed through a time-distributed dense layer to generate the final audio feature predictions. The dense layer has 80 units, corresponding to the dimension of the target audio features. A linear activation function is used in the output layer since the target audio features are continuous values. We train the spectrogram generation model on 20 audible samples of each of the phrases, isolated commands, and digits per user, using custom loss functions as discussed in §5.3.

5.3 Custom Loss Functions

We introduce a custom loss function that combines 6 metrics to ensure that the generated spectrogram closely matches the target spectrogram.

• **The weighted spectral convergence loss** ($\mathcal{L}_{WSC}$) is calculated as the weighted mean squared error (MSE) between the predicted spectrogram $\hat{S}(f,t)$ and the target spectrogram $S(f,t)$, where f and t represent the frequency and time dimensions, respectively. The weights $W(f)$ are assigned to each frequency bin to emphasize the importance of lower frequency ranges, which often contain more critical speech information [9, 79].

$$\mathcal{L}_{WSC} = \frac{1}{FT} \sum f = 1^{F} \sum_{t=1}^{T} W(f)(\hat{S}(f,t) - S(f,t))^2 \quad (1)$$

$W(f)$ is the masking function that assigns weights to each frequency bin based on the linear increase from 0 to 4000 Hz with a window of 100 Hz.

• To ensure that the generated speech accurately captures the phonetic content of the input text, we introduce a **phoneme-level loss**($\mathcal{L}_{PL}$) that compares the predicted phoneme sequence with the target phoneme sequence. This loss is calculated using Connectionist Temporal Classification (CTC):

$$\mathcal{L}_{PL} = -\sum_{t=1}^{T} \log p(\pi_t | \mathbf{y}_t) \quad (2)$$

where π_t represents the target phoneme at time step t, and $\mathbf{y}_t$ represents the predicted phoneme probability distribution.

• We incorporate a **correlation loss** ($\mathcal{L}_{Corr}$) between the audio time-domain envelope and the predicted time-domain envelope. The predicted time-domain envelope is obtained by applying an inverse short-time Fourier transform (ISTFT) to the generated spectrogram and then creating an envelope. This loss encourages the model to learn features from the temporal audio domain, such as the number of syllables:

$$\mathcal{L}_{Corr} = 1 - \frac{\sum_{t=1}^{T}(x(t) - \bar{x})(\hat{x}(t) - \bar{\hat{x}})}{\sqrt{\sum_{t=1}^{T}(x(t) - \bar{x})^2}\sqrt{\sum_{t=1}^{T}(\hat{x}(t) - \bar{\hat{x}})^2}} \quad (3)$$

where $x(t)$ and $\hat{x}(t)$ represent the audio time-domain envelope and the predicted time-domain envelope, respectively, and $\bar{x}$ and $\bar{\hat{x}}$ represent their respective means.

• To generate speech with more natural prosody (i.e., rhythm, stress, and intonation), we incorporate a **prosody loss** ($\mathcal{L}_P$) that compares the predicted pitch and energy contours with the target contours. This loss is calculated using dynamic time warping (DTW):

$$\mathcal{L}_P = DTW(\hat{C}, C) \quad (4)$$

where $\hat{C}$ and C represent the predicted and target pitch and energy contours, respectively.

• We use **cosine similarity loss** ($\mathcal{L}_{CS}$) between the model's higher dimensional representation and the text embeddings. The higher dimensional representation is obtained by aggregating the outputs of a dense layer in the model. By minimizing the cosine distance between these representations, the model learns to generate similar spectrograms for IMU signals corresponding to the same word:

$$\mathcal{L}_{CS} = 1 - \frac{\mathbf{h} \cdot \mathbf{e}}{\|\mathbf{h}\|\|\mathbf{e}\|} \quad (5)$$

where $\mathbf{h}$ represents the model's higher dimensional representation and $\mathbf{e}$ represents the text embedding.

• To account for the continuous nature of speech and allow for abrupt changes around the word boundaries, we introduce a **temporal coherence loss**($\mathcal{L}_{TC}$). This loss is calculated by applying a masking layer to the spectral convergence loss, with values of 0 near the word boundaries and 1 elsewhere:

$$\mathcal{L}_{TC} = \frac{1}{FT} \sum_{f=1}^{F} \sum_{t=1}^{T} M(t)(\hat{S}(f,t) - S(f,t))^2 \quad (6)$$

where $M(t)$ represents the masking function.

To ensure that each loss function contributes equally to the total loss, we normalize the individual losses. The weights for each loss function are determined using a grid search and the total loss is then calculated as a weighted sum of the normalized individual loss functions. During training, the model weights are optimized to minimize the total loss using standard optimization techniques. By combining these carefully selected loss functions, the model learns

to generate accurate spectrograms from jaw motion data. Figure 5 shows the effect of skipping different loss functions on spectrogram generation.

6 LLM-assisted text generation

After training the model, the final step in the inference phase is to convert the spectrogram into text. It's important to note that while we utilize additional inputs like GPT-2 text tokens and audio features during training to enhance the model's learning process, these are incorporated into the loss functions and are not required during inference. At inference time, our model generates spectrograms solely from IMU data.

For speech recognition using spectrogram instead of developing a custom model, we use Whisper [72] speech recognition model. To construct phrases from the generated word spectrogram, we stitch together the spectrogram for each word, inserting 0.2 seconds of silence between words. Recall from Section 4 that some of the words were eliminated from the input. These words may contain important information about the phrase and need to be filled into the final sentence. To fill in the missing words (blanks) for speech-based interfaces language-based models have been widely adopted in previous works [38, 96, 98]. While rule-based or template-based language models could potentially fill these gaps, they often fail in critical cases. For instance, if the word "add" is skipped from the phrase "Add apples to shopping list", a simple model might incorrectly insert "remove" based on its training data, completely altering the intended meaning. To overcome this limitation of predictive language models, we leverage generative models which can fill in the blanks with multiple candidates, and the most appropriate can be chosen based on the application context. We use a general LLM (GPT4) instead of ASR-specific language models like Kaldi [69] as it provides greater flexibility in handling ambiguous or partial inputs, which is crucial in a speech interface [56]. While many language models like BERT exist, we specifically chose GPT for text completion and correction. GPT is a state-of-the-art language model that allows us to use prediction probabilities in the form of prompts without the need for any fine-tuning or training. While similar capabilities might be possible with BERT, it would require complex fine-tuning as it is trained on text data from sources like Wikipedia, unlike GPT which is trained on instruction sets [74]. This characteristic of GPT aligns well with our need for flexible and context-aware text generation.

We carefully craft input prompts that encourage GPT to produce outputs with contrastive intents, ensuring a range of contextually appropriate options. However, it is not straightforward to integrate GPT with our system. This is because while GPT can generate multiple candidates for the blanks, there is not enough information available in the skipped word segments to determine the correct candidate. To refine GPT's output, we utilize linguistic information from the skipped words. While we cannot extract all phonetic components from these segments, we can still derive valuable data, such as the number of syllables and the vowel category (low, mid, or high) [8]. High vowels (like 'i' in "eat" or 'u' in "boot") require minimal jaw opening, with the mouth remaining relatively closed. Mid vowels (like 'e' in "bet" or 'o' in "boat") involve a moderate jaw opening. Low vowels (like 'a' in "father") necessitate the widest

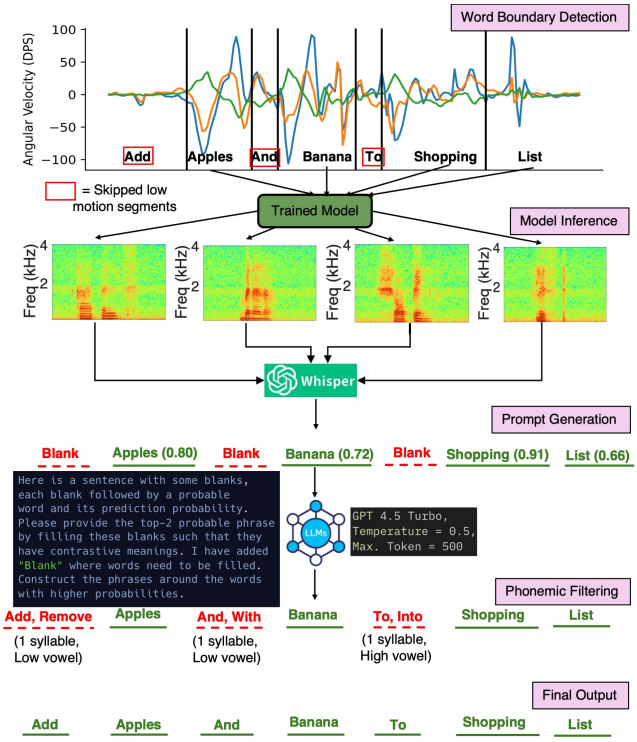


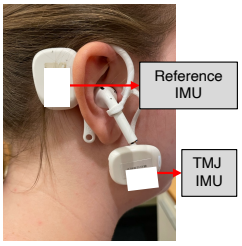
Figure 6: Steps in *Unvoiced* for the phrase *Add apples and bananas to shopping list.*

jaw opening, with the mouth in its most open position. These distinctions in jaw motion contribute to the unique jaw motion properties of each vowel category [50]. We extract 4 statistical features: mean, area under the curve, skewness, kurtosis, and the first 8 FFT coefficients and train an SVM for vowel category classification. This classifier processes the linguistic information to identify the best candidate among GPT's suggestions, effectively filling in the missing words. Our approach allows us to incorporate GPT and potentially other large language models without the need for fine-tuning, while still maintaining context-appropriate and linguistically plausible outputs. Figure 6 shows our GPT prompt and end-to-end inference pipeline. Whisper's predictions, along with their probabilities and a "Blank" token, are input to GPT. As we do not expect GPT to fill long words or generate creative text we set a small max token length of 500 and temperature to 0.5 (controls the randomness). Finally, we filter the GPT output using phonetic information from "Blank" token word segments to produce the final text.

7 Implementation and Setup

This section will explain our user study design, metrics, and baselines for evaluation.

■ **Data Collection** Our earable prototype incorporates a twin-IMU setup to mitigate body motion artifacts for jaw motion [83], strategically placing one IMU on the TMJ and the other on the temporal bone. We log timestamps, 3-axis accelerometer (± 2g), 3-axis gyroscope (250 DPS) data, and 48kHz audio. We conducted an



Setup	Samples	Sessions	Users
Interaction tasks	5 phrases/tasks, 20 U/phrase	1	19
Phrases	30 phrases, 20 U, 20 V	1	19
Isolated Commands and Digits	20 commands, 0-9 digits 20 U, 20 V	1	19
Numerical Data	10 (6 digits), 10 U	1	8
Motion and Acoustic Noise	5 phrases/tasks, 10 U/phrase	1	8

Figure 7: Our prototype is inspired by [83]. Our system was evaluated in different settings: U-Unvoice, V-voiced.

Task	Example Phrases
Media Control (Task 1)	Play music on YouTube, Play songs from dance playlist.
Home Control (Task 2)	Order the speaker from wishlist, Set alarm for 6 AM.
Messaging (Task 3)	Send: On my way home talk soon, Message mom talk to you later.
Quick Notes (Task 4)	Add Note: Buy apples on the way back home, Save note: Meeting at 3 PM with the advisor.

Table 1: We collect data for 4 common voiced tasks.

IRB-approved user study with 19 participants (11 females, 8 males) in noise-free and different noisy conditions. All our participants were fluent in English and had varied native languages, including Hindi (6), English (5), Tamil (3), Spanish (3), and Kannada (2) which challenges the system's ability to process various accents, dialects, and pronunciation patterns. We collected data in three settings: (1) free of motion and acoustic noise, (2) with motion noise, and (3) with acoustic noise. Data collection occurred in a quiet room, and users were instructed to sit still without performing body movements unless otherwise stated. The user study aimed to assess (a) the system's ability to achieve the user's expected output, (b) its accuracy for phrases, numerical codes, and words, (c) the precision of spectrograms, and (d) robustness compared to a voiced assistant. Our data collection prototype and summary are shown in Figure 7. *Interaction tasks (Row 1):* We evaluate the system's performance as an UUI for four tasks shown in Table 1. Each task consisted of 5 phrases, repeated 20 times unvoiced reflecting scenarios when voice-based systems are affected by noise, or the user requires private communication. *Phrases (Row 2):* We asked participants to speak 30 phrases, 20 voiced, 20 unvoiced. These phrases ranging from 2 to 11 words are commonly used for interaction with voice assistants and do not overlap with the interaction phrases. *Isolated Commands and Digits (Row 3):* We selected twenty single or double-word commands with multiple syllables. We instructed users to repeat each command and digit (0-9) 20 times voiced, 20 unvoiced. *Numerical Data (Row 4):* We asked users to create 10 unique 6-digit passcodes and repeat each passcode 10 times unvoiced. *Noisy Conditions (Row 5):* We asked 8 users to repeat each phrase in the 4 tasks 10 times unvoiced. We collected data (a) seated and moving freely, performing head motions, (b) walking in a corridor, and (c) with restaurant and machine noise (60dB) playing in the background.

■ **Metrics.** We use the following metrics.

• *Task Completion Rate (TCR)*: TCR is the proportion of times that the system generates the expected output for a given silent phrase (max/best value: 1). For media and smart home control, we convert outputs to speech for a voice assistant, assigning TCR 1 if executed correctly and 0 otherwise. For messaging and note-taking, we compare outputs word-by-word with inputs. Perfect matches receive TCR 1; even if the output is off by a single word, TCR is 0.

• *Word Error Rate (WER)*: WER quantifies the minimum number of edit operations (substitutions (S), deletions (D), and insertions (I)) needed to transform the predicted text into the ground truth, normalized by the total number of words (N) in the ground truth [51, 105].

• *Top-1 Accuracy*: Top-1 accuracy measures the fraction of test instances for which the correct label is among the top predicted labels, respectively, when the model outputs a ranked list of predictions [83, 96].

• *Spectral Convergence (SC)*: Spectral Convergence is our primary metric for evaluating the quality of reconstructed spectrograms. SC quantifies the similarity between the reconstructed spectrogram $\hat{S}(f,t)$ and the target spectrogram $S(f,t)$ in the frequency domain [34, 93]. A SC score of 0 indicates perfect reconstruction, with no difference between the reconstructed and ground truth spectrograms. Unlike previous works [19] that employ metrics such as Short-term objective intelligibility [88] or Perceptual evaluation of speech quality [76], our focus is on generating accurate spectral features rather than optimizing for speech signal intelligibility or quality. This approach aligns with our goal of producing high-fidelity spectrograms for subsequent processing, rather than directly generating audible speech signals.

■ **Baselines**. We compare against the following baselines. It's important to note that our LLM-assisted approach is an integral part of *Unvoiced*'s system design, not an additional feature. The prompt engineering, which includes specifying blank locations and vowel categories derived from our algorithm, is a key component of our method. This approach demonstrates the potential of combining traditional techniques with LLM assistance. While the baselines represent standard implementations and do not use LLMs, *Unvoiced* showcases how LLM integration can enhance performance in unvoiced speech recognition tasks.

• *MuteIt* (Signal Processing Based): MuteIt [83] is designed for silent speech recognition of isolated single-word commands spoken slowly (at a lower than normal speech rate). To match the word-level input, we apply our phrase segmentation technique to extract individual words first. We utilize MuteIt's technique to extract vowels and other phonetic components, training a personalized model for each user with the same amount of data as suggested by the authors.

• *LIMU-BERT* (IMU Modelling Based): We use their pre-trained model [99] to extract representations, followed by a bidirectional LSTM head serving as a classifier. Similar to MuteIt, we train a personalized model for each user.

• *BiDiLSTM*(Deep Learning Based): Lastly, we implement a Recurrent Neural Network (RNN) model called BiDiLSTM, which consists of two Bidirectional LSTM layers, five fully connected layers, and a classifier layer.

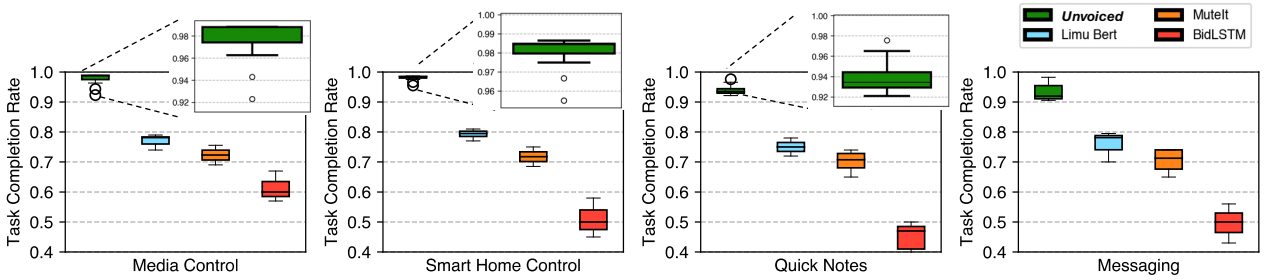


Figure 8: The Mean TCR for our system is more than 93% for 4 different tasks. ↑ is better.

For LIMU-BERT and BiDiLSTM, we construct an n-class classifier for n words in the dictionary and perform an 80-20 train-test split on our data. These models are fine-tuned with 15% of the data from the train set to ensure a fair comparison by fine-tuning the baseline models for optimal performance.

8 Evaluation

In this section, we present details on the 5 key takeaways from our evaluation. (1) *Unvoiced* achieves a task completion rate exceeding 94% while maintaining a word error rate below 0.09. (2) It can be seamlessly integrated with existing off-the-shelf speech recognition and large language models (LLMs) without requiring any fine-tuning. (3) We showcase the system's ability to generate high-fidelity spectrograms for words outside the training data. (4) Our system exhibits robustness to motion noise and surpasses Siri's task completion rate in the presence of acoustic noise, demonstrating its resilience in challenging environments. Finally, our exit survey reveals that 90% of users found our system comfortable and easy to use.

■ ***Unvoiced* as a User Interface.** We report the Task Completion Rate (TCR) of *Unvoiced* compared to the three baseline methods across four different tasks in Figure 8. Our system achieves TCR of 98.6%, 97.7%, 94.8%, and 93.1% for tasks 1, 2, 3, and 4, respectively, and demonstrates a substantial improvement of at least 26% over any of the baseline methods for all tasks. Notably, our system exhibits higher TCR for tasks 1-2 than for tasks 3-4. This can be attributed to the inherent differences in these tasks. Tasks 1-2 are evaluated using Siri, which operates within a more limited context. In contrast, tasks 3-4 involve text editing, which requires a lower Word Error Rate (WER) to be considered complete. These results validate our system's efficacy in achieving the user's desired output for various applications.

■ **Continuous Speech Recognition**. We present the Word Error Rate (WER) for recognizing phrases and numerical passcodes to show the ability of continuous speech recognition in Figure 9. Our approach yields a WER of less than 10% for over 90% of the phrases, a 35% reduction in WER compared to any of the baseline methods. These results demonstrate that the spectrograms generated by our system can be successfully fed into an off-the-shelf speech recognition model without requiring any fine-tuning, highlighting the compatibility of our approach.

■ **Isolated Command Recognition.** We depict the top-1 accuracy of recognizing isolated commands, where individual commands

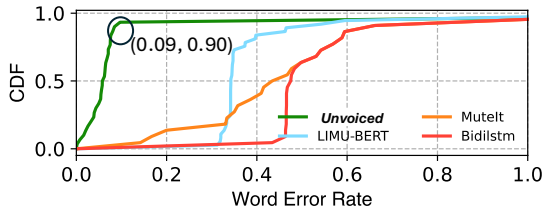


Figure 9: Word Error Rate (WER) for our system is as low as 9% for more than 90% of the phrases. ↓ is better.

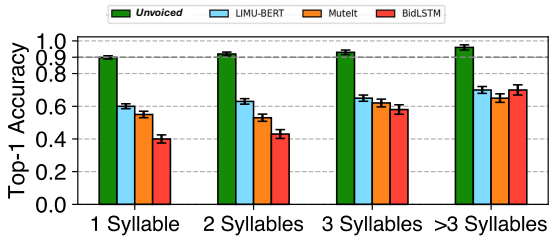


Figure 10: We achieve more than 90% Top-1 for words of different syllabic lengths. ↑ is better.

are issued without the context of surrounding words in Figure 10 for mono-, di-, tri-, and multi-syllabic (>3 syllables) words. We surpass 90% accuracy for all word categories, highlighting our ability to effectively recognize unvoiced speech even in the absence of contextual information. In contrast, the baseline methods fail to achieve accuracy scores above 72%. Even MuteIt, a system designed for isolated command recognition, does not perform well in this evaluation with a normal speech rate.

■ **Spectrogram Generation.** It is important to evaluate how our audio spectrogram generation preserves important speech information. Figure 11(a) illustrates the audio spectrogram for the phrase "Read notification from bedroom speaker" alongside the corresponding IMU-generated spectrogram Figure 11(b). We observe similarity in (1) energy configuration, (2) pitch variation patterns (start and end of different energy events within the spectrogram), (3) temporal alignment, and the correct number of syllables. This indicates the system's ability to preserve intonation and speech rate information and verifies the importance of employing carefully crafted loss functions.

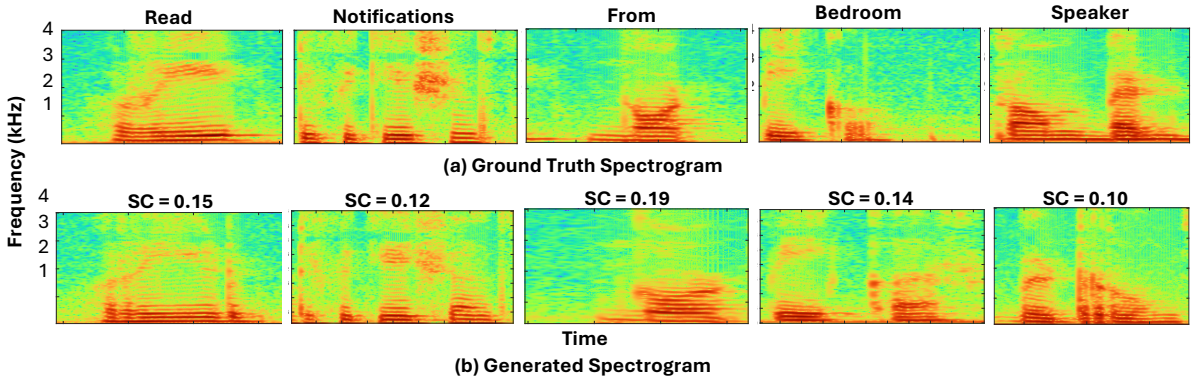


Figure 11: Spectrogram for the phrase "Read notification from bedroom speaker". Our system is able to achieve low spectral convergence (SC).

	Spectral Convergence	RMSE (x10-5)	WER
In Training Phrases	0.14 (0.12)	0.24 (0.21)	0.07±0.008
Out of Training Phrases	0.16 (0.13)	0.28 (0.25)	0.11±0.006
Unseen Words	0.22 (0.18)	0.32 (0.27)	0.14±0.009

Table 2: Performance Metrics: SC, RMSE, and WER for training schemes. Parenthesis contain SC and RMSE for the same phrase in another session from audio data. WER includes mean and SD. ↓ SC, RMSE, and WER are better.

Figure 12: Our system is robust to different types of motion and acoustic noises. ↑ is better.

To quantitatively assess the system's spectrogram generation capabilities, we utilize the Spectral Convergence (SC) metric. Table 2 presents the SC and Root Mean Square Error (RMSE) values for different data split techniques. Considering inherent variability between word samples and the fact that the generated spectrogram relies solely on jaw motion data, *Unvoiced* achieves SC scores below 0.17, showing the system's strong performance in generating accurate spectrograms.

■ **Unseen Words Recognition.** To further investigate whether the system is learning a one-to-one mapping between IMU data and audio, we conduct tests on words that are subsets of the training words but have never been encountered by the model before. We make a subset of 50 unseen words from the dataset and report low SC and WER <0.15 (Table 2). These findings provide evidence of *Unvoiced*'s ability to generate high-fidelity spectrograms that are close to the spectral-temporal features of ground truth, even for words and phrases outside the training set.

■ **Impact of Noise.** We analyze the robustness of *Unvoiced* under different motions (head motion and walking) and acoustic noise (machine and restaurant).

Motion Noise: We evaluate the impact of motion noise on our system's performance by training the model using noise-free data and testing on data with motion noise. The first two bars in Figure 12 present the mean Task Completion Rate (TCR) for head motion and walking. The high TCR (>0.9) indicates that motion noise resulting from body movement has minimal effect on the system's performance.

Acoustic Noise: A key advantage of our system is its ability to function effectively in noisy acoustic conditions. To test this, we play two common noise types, machine noise (drilling) and people talking in a restaurant (cocktail party), at 60 dB. Since our model does not rely on acoustic features during inference, we achieve a high TCR of 93% in both noisy conditions, Figure 12 last two bar sets.

We compare our system's performance with Siri in both noisy environments, not as a speech recognition system but as an interaction medium. We use Siri as a baseline since it is one of the most common speech-based interaction systems available on commercial mobile devices. While Siri outperforms our system by 13% in terms of TCR in the presence of motion noise, our system achieves a higher TCR when exposed to acoustic noise. This suggests that our system can be a preferred choice for interaction over traditional voice assistants when moving through noisy environments, such as hallways.

■ **Impact of Different Factors.** In this section, we gauge the effect of different settings (training data, phrase length and rate, native language, wearing position, and choice of speech recognition model) on the system's performance.

Impact of Speech rate and Length. User-dependent features, such as speech rate and content length, can significantly influence our system's performance. Figure 13 shows the effect of speech length and rate on the Word Error Rate (WER) for phrases. We evaluate speech rate at multiple levels (<50, 50-100, 100-150, >150 WPM). WER decreases as the number of words increases, aligning with the expectation that the Whisper model relies on contextual information. Our system maintains a WER lower than 0.1 when

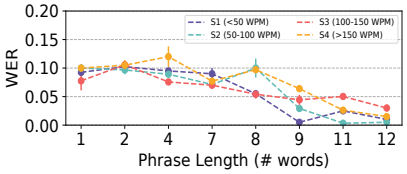


Figure 13: WER for phrases with various lengths and speech rates. Even with the phrase's short length (4 words) and high speech rates (>150 WPM), our WER was < 0.15.

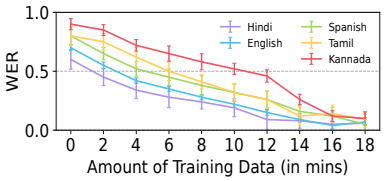


Figure 14: We achieve WER <0.13 with only 16 minutes of training data for a user pool of 5 different native languages.

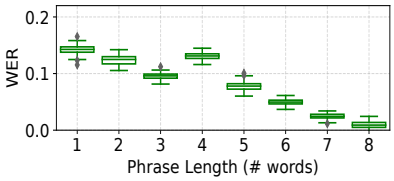


Figure 15: Out-of-training performance: WER is 0.15 for 1-word commands and as low as 0.07 for 5-word phrases.

the number of words exceeds 7. Remarkably, even at fast speech rates above 150 WPM, our model's WER never exceeds 0.12, showcasing its ability to capture rapid changes in speech content. The above results show that our system can maintain high robustness at various speech rates, especially in medium and long sentences.

Impact of Training Size. We train a personalized model for each user. However, to make the system widely acceptable minimal data must be needed from a new user. Toward this end, we investigate the data required from each user to achieve WERs. As previous works have shown it is efficient to train a native language-specific model [14, 92], we categorize users based on native languages and perform leave-one-user-out evaluation. We fine-tuned the model for the test user and achieved 0.12 WER with only 15 minutes of data for 4 out of 5 languages (Figure 14). The Kannada language requires more data, which can be obtained from a single Kannada-speaking user. As we expand our user pool and employ domain adaptation techniques, we expect to further reduce the WER while requiring less data. These findings highlight the adaptability and scalability of our system, as it effectively learns from limited user-specific data to accommodate linguistic variations with minimal data requirements.

Out of training phrases. We evaluate our system's ability to recognize phrases that are not included in the training set. The words in these phrases may be present in the training at a different location. For each user, we randomly select 15 phrases for training and use the remaining phrases for testing, repeating the experiment until all phrases have been tested at least once. The average WER across all users and phrases is 0.14. Figure 15 illustrates the WER for different sentence lengths. We observe a clear decreasing trend in WER as the sentence length increases. For single-word sentences, the WER is 0.14 and decreases to 0.08 when the sentence contains seven words. This behavior is expected, as longer sentences provide more contextual information that can be effectively captured by the end-to-end network, leading to improved recognition accuracy. These results demonstrate our system's ability to generalize well to unseen sentences, even with limited training data per user.

Longitudinal Study. To assess our system's performance in capturing speech mannerisms over time as well as robustness to variations in sensor placement, we conducted a 25-day longitudinal study with 8 users. We utilized data collected during the initial training phase and randomly selected 20 phrases. Users were asked to silently articulate these phrases 10 times each in multiple sessions over 3 weeks, resulting in an average of 8 sessions and 160 samples per user. Word Error Rate (WER) was consistently below 0.15 throughout the 4-week period, confirming the robustness and stability of our system over time. The low WER highlights the system's adaptability to individual speech patterns and its capacity to preserve learned mannerisms.

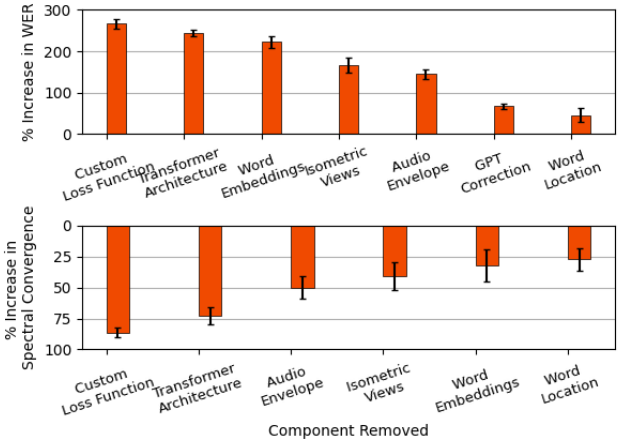


Figure 16: Increase in WER and SC when removing system components. ↓ WER and SC are better.

■ **Ablation Study.** We conduct ablation experiments to quantitatively investigate the performance of spectrogram generation. We remove each component of the system one by one and report the increase in WER and increase in SC. Figure 16 shows our findings.

Without Custom Loss: We do not use our loss function. Instead, we use mean squared error as the loss function. Removing the custom loss function causes the highest increase in the WER of the system.

Without Transformer Architecture: We do not use multi-head attention mechanisms in our system. Instead, we use a traditional RNN network with no attention blocks, causing an increase of almost 250% in WER.

Without Isometric Views: We input 6-axis IMU data into the model after removing the effect of gravity and orientation from the accelerometer and normalizing the 6-axis data. This causes a 73% increase in the spectral convergence.

Without word embedding and audio time domain features: In the model training, we skip the block of correlating word embeddings with the input IMU data and the audio time domain features

like the number of syllables, signal envelope, and other audio time-domain features. The removal of these blocks increases the WER by 178% and 68% respectively.

Without phonetic correction: In this experiment, we do not correct the output from the GPT but directly use the most probable phrase as the prediction.

9 Related Work

Unvoiced User Interfaces have garnered significant attention as an alternative mode of interaction, particularly for their non-intrusive nature and high user acceptance. While UUIs have recently been explored extensively, UUIs with LLMs are still uncharted space.

VUI with LLMs. VUIs have gained significant attention in recent years, leveraging the advancements in speech recognition and language modeling [29]. LLMs have been employed to enhance the natural language understanding and generation capabilities of voice-based systems [100]. They can capture the context and semantics of spoken queries, enabling more accurate and contextually relevant responses [2, 73, 110]. Additionally, they have been used to generate human-like responses, improving the naturalness of the generated speech [27].

Contact Free Methods. Contact-free UUIs leverage non-physical touch sensors, like cameras, wifi, and ultrasonic sensors. Early implementations mainly employed camera-based technologies to capture facial movements, particularly lip motions, using extensive video datasets to achieve robust performance despite challenges such as variable lighting and privacy concerns [26, 62, 63, 67, 86, 97, 102] Technologies like Radio Frequency (RF) signals [18, 82], and ultrasonic [10, 13, 44] have also been employed to track articulator movements. While these systems address the privacy issues of camera-based systems, they are sensitive to environmental conditions, require calibration, and can be affected by interference from other devices, which impacts their robustness. Recent advancements have shifted towards acoustic sensing technologies, which utilize high-frequency sound transmissions from mobile devices captured by microphones to infer articulator movements [15, 19, 22, 32, 33, 51, 89, 105, 109]. Although these methods are portable and less affected by environmental conditions, they often require user interaction, such as holding the device, which can be impractical in situations like driving or for individuals with accessibility needs. Our system provides a hands-free and robust alternative to voice-based interactions.

Wearable Methods. Wearable UUIs typically involve placing one or more sensors on the face, inside the mouth, or on the articulators to detect unvoiced speech. Electromyography (EMG) sensors are used to measure muscle activity in the lips, jaw, and cheeks during speech production [37, 55]. These sensors, which require attaching skin electrodes around the cheek and lips [21, 57] are often not socially acceptable and difficult to integrate with commercial wearable products. Sensors placed on the articulators [23, 30, 43, 48, 75, 81, 95], or even retrofitted to masks [28], can capture articulator motion and infer unvoiced speech. However, many of these techniques are intrusive, involving magnetic sensors on the tongue or inside the mouth, or tattoos around the lips, making them socially unacceptable and requiring calibration. IMUs placed on the temporomandibular joint (TMJ) have been used to capture jaw movements during speech production, but these systems can only recognize phonemes, or words, limiting their applications [39, 83, 85, 106]. While speech enhancement systems offer a viable method for speech-based interactions in noisy environments, they require users to vocalize commands, which might not be desirable for private information sharing or discreet communication. Moreover, these systems often utilize multiple sensory inputs (IMU and microphone [29, 97] or ultrasound and microphone [14, 16, 88]) at inference, unlike our system which only needs IMU input, making it more challenging. With *Unvoiced*, we present, to the best of our knowledge, the first ear-worn UUI that can perform multiple interaction tasks in both noise-free and noisy conditions in an unvoiced manner using only an IMU. This approach overcomes the limitations of previous methods, offering a more socially acceptable and versatile solution for unvoiced speech interaction.

10 Discussion

Unvoiced is the first jaw motion-based UUI that can enable silent communication with devices that expect voice input. It is a first step in making unvoiced interactions as common as voice interactions. There are, however, areas for improvement that we aim to address in the future. (1) Currently, we train personalized jaw motion to spectrogram models for each user. We will investigate techniques that require minimal training data from each user and explore possibilities for a generalized model. (2) We utilize an early research prototype. While the position of sensors is similar to many commercially available earables, we will assess compatibility with a range of commercial devices. (3) The system evaluation covers limited types of noise. In the future, we will evaluate performance in the presence of other motion sources, such as traveling in a car, running, etc. (4) We plan to conduct a more robust study with a larger, more diverse user base to better understand the effect of factors such as native language and vocabulary size on system performance. (5) The current study did not specifically address age-related variability or users with speech impediments. In the future [1], we will include participants from these groups to ensure the system's effectiveness across a broader range of users.

11 Conclusion

In conclusion, *Unvoiced* presents a novel approach to silent device interaction using earables. At its core, *Unvoiced* translates jaw motion, captured via IMU, into Mel spectrograms. This cross-modal translation incorporates phonetic, contextual, and syntactic information, while our specialized loss function optimizes for these linguistic features, ensuring the generated spectrograms capture nuanced speech characteristics. By exploiting recent advances in LLMs and combining them with our spectrogram reconstruction technique, *Unvoiced* achieves a remarkable >94% task completion rate for 4 diverse applications that typically expect voice input.

Acknowledgments

We thank the shepherd and the anonymous reviewers for their insightful comments. This material is based in part upon work supported by the National Science Foundation under Award Numbers 2238553, 2401415, and 2403528, and by the Hellman Fellows Award.

References

[1] 2024. Enabling Accessible and Ubiquitous Interaction in Next-Generation Wearables: An Unvoiced Speech Approach. In *The 30th Annual International Conference on Mobile Computing and Networking* (Washington D.C., DC, USA) *(ACM MobiCom '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3636534.3695908

[2] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977* (2020).

[3] Amazon. 2024. Alexa. https://developer.amazon.com/en-US/alexa

[4] Amazon. 2024. Alexa Home Assistant. https://www.home-assistant.io/integrations/alexa/

[5] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, et al. 2015. Deep speech 2: End-to-end speech recognition in english and mandarin. CoRR abs/1512.02595 (2015). *arXiv preprint arXiv:1512.02595* (2015).

[6] Apple. 2024. Siri. https://www.apple.com/siri/

[7] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech communication* 56 (2014), 85–100.

[8] Encyclopedia Britannica. 2024. Vowel. https://www.britannica.com/topic/vowel

[9] Gregory B Cogan and David Poeppel. 2011. A mutual information analysis of neural coding of speech by low-frequency MEG phase information. *Journal of neurophysiology* 106, 2 (2011), 554–563.

[10] Tamás Gábor Csapó, Csaba Zainkó, László Tóth, Gábor Gosztolya, and Alexandra Markó. 2020. Ultrasound-based articulatory-to-acoustic mapping with WaveGlow speech synthesis. *arXiv preprint arXiv:2008.03152* (2020).

[11] Saeed Dabbaghchian, Marc Arnela, Olov Engwall, and Oriol Guasch. 2019. Reconstruction of vocal tract geometries from biomechanical simulations. *International journal for numerical methods in biomedical engineering* 35, 2 (2019), e3159.

[12] Dataquest. 2024. What are Large Language Models (LLMs) and how will they be used in 2024? https://www.dataquest.io/blog/what-are-large-language-models-llms-and-how-will-they-be-used-in-2024/

[13] B. Denby, Y. Oussar, G. Dreyfus, and M. Stone. 2006. Prospects for a Silent Speech Interface using Ultrasound Imaging. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1. I–I. https://doi.org/10.1109/ICASSP.2006.1660033

[14] Li Deng and Xiao Li. 2013. Machine Learning Paradigms for Speech Recognition: An Overview. *IEEE Transactions on Audio, Speech, and Language Processing* 21, 5 (2013), 1060–1089. https://doi.org/10.1109/TASL.2013.2244083

[15] Xuefu Dong, Yifei Chen, Yuuki Nishiyama, Kaoru Sezaki, Yuntao Wang, Ken Christofferson, and Alex Mariakakis. 2024. ReHEarSSE: Recognizing Hidden-in-the-Ear Silently Spelled Expressions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[16] Ecobee. 2024. Ecobee Camera. https://www.ecobee.com/en-us/cameras/smart-camera-with-voice-control/

[17] Donna Erickson. 2002. Articulation of extreme formant patterns for emphasized vowels. *Phonetica* 59, 2-3 (2002), 134–149.

[18] David Ferreira, Samuel Silva, Francisco Curado, and António Teixeira. 2022. Exploring silent speech interfaces based on frequency-modulated continuous-wave radar. *Sensors* 22, 2 (2022), 649.

[19] Yongjian Fu, Shuning Wang, Linghui Zhong, Lili Chen, Ju Ren, and Yaoxue Zhang. 2023. SVoice: Enabling Voice Communication in Silence via Acoustic Sensing on Commodity Devices. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems* (<conf-loc>, <city>Boston</city>, <state>Massachusetts</state>, </conf-loc>) *(SenSys '22)*. Association for Computing Machinery, New York, NY, USA, 622–636. https://doi.org/10.1145/3560905.3568530

[20] Yongjian Fu, Shuning Wang, Linghui Zhong, Lili Chen, Ju Ren, and Yaoxue Zhang. 2024. UltraSR: Silent Speech Reconstruction via Acoustic Sensing. *IEEE Transactions on Mobile Computing* (2024), 1–18. https://doi.org/10.1109/TMC.2024.3419170

[21] David Gaddy and Dan Klein. 2020. Digital Voicing of Silent Speech. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 5521–5530. https://doi.org/10.18653/v1/2020.emnlp-main.445

[22] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-Based Silent Speech Interface for Smartphone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 80 (Sept. 2020), 27 pages. https://doi.org/10.1145/3411830

[23] Jose A. Gonzalez, Lam A. Cheah, Angel M. Gomez, Phil D. Green, James M. Gilbert, Stephen R. Ell, Roger K. Moore, and Ed Holdsworth. 2017. Direct Speech Reconstruction From Articulatory Sensor Data by Machine Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 12 (2017), 2362–2374. https://doi.org/10.1109/TASLP.2017.2757263

[24] Jose A Gonzalez-Lopez, Alejandro Gomez-Alanis, Juan M Martín Doñas, José L Pérez-Córdoba, and Angel M Gomez. 2020. Silent speech interfaces for speech restoration: A review. *IEEE access* 8 (2020), 177995–178021.

[25] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints* (2023).

[26] J. Han, L. Shao, D. Xu, and J. Shotton. 2013. Enhanced Computer Vision With Microsoft Kinect Sensor: A Review. *IEEE Transactions on Cybernetics* 43, 5 (2013), 1318–1334. https://doi.org/10.1109/TCYB.2013.2265378

[27] Hongkun Hao, Long Zhou, Shujie Liu, Jinyu Li, Shujie Hu, Rui Wang, and Furu Wei. 2023. Boosting large language model for speech synthesis: An empirical study. *arXiv preprint arXiv:2401.00246* (2023).

[28] Hirotaka Hiraki and Jun Rekimoto. 2021. SilentMask: Mask-Type Silent Speech Interface with Measurement of Mouth Movement. In *Augmented Humans Conference 2021* (Rovaniemi, Finland) *(AHs'21)*. Association for Computing Machinery, New York, NY, USA, 86–90. https://doi.org/10.1145/3458709.3458985

[29] Julia Hirschberg and Christopher D Manning. 2015. Advances in natural language processing. *Science* 349, 6245 (2015), 261–266.

[30] 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. https://doi.org/10.1016/j.specom.2012.02.001

[31] Hajar Homayouni, Sudipto Ghosh, Indrakshi Ray, Shlok Gondalia, Jerry Duggan, and Michael G Kahn. 2020. An autocorrelation-based LSTM-autoencoder for anomaly detection on time-series data. In *2020 IEEE international conference on big data (big data)*. IEEE, 5068–5077.

[32] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone. 2010. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication* 52, 4 (2010), 288–300.

[33] Thomas Hueber, Gérard Chollet, Bruce Denby, and Maureen Stone. 2008. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. *Proc. of ISSP* (2008), 365–369.

[34] Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. 2021. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. *arXiv preprint arXiv:2106.07889* (2021).

[35] Frederick Jelinek. 1998. *Statistical methods for speech recognition*. MIT press.

[36] Daniel Jurafsky and James H Martin. [n. d.]. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

[37] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. Alterego: A personalized wearable silent speech interface. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*. 43–53.

[38] Himmet Toprak Kesgin and Mehmet Fatih Amasyali. 2023. Iterative mask filling: An effective text augmentation method using masked language modeling. In *International Conference on Advanced Engineering, Technology and Applications*. Springer, 450–463.

[39] Prerna Khanna, Tanmay Srivastava, Shijia Pan, Shubham Jain, and Phuc Nguyen. 2021. JawSense: recognizing unvoiced sound using a low-cost ear-worn system. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*. 44–49.

[40] Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. How Well Do LLMs Represent Values Across Cultures? Empirical Analysis of LLM Responses Based on Hofstede Cultural Dimensions. *arXiv preprint arXiv:2406.14805* (2024).

[41] Kia. 2024. Kia Voice Control. http://webmanual.kia.com/STD_GEN5_WIDE/AVNT/EU/English/voicerecognitionsystem.html

[42] Sungwon Kim, Kevin Shih, rohan badlani, Joao Felipe Santos, Evelina Bakhturina, Mikyas Desta, Rafael Valle, Sungroh Yoon, and Bryan Catanzaro. 2023. P-Flow: A Fast and Data-Efficient Zero-Shot TTS through Speech Prompting. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 74213–74228. https://proceedings.neurips.cc/paper_files/paper/2023/file/eb0965da1d2cb3fbbbb8dbbad5fa0bfc-Paper-Conference.pdf

[43] Naoki Kimura, Tan Gemicioglu, Jonathan Womack, Richard Li, Yuhui Zhao, Abdelkareem Bedri, Zixiong Su, Alex Olwal, Jun Rekimoto, and Thad Starner. 2022. SilentSpeller: Towards mobile, hands-free, silent speech text entry using electropalatography. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[44] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. In *Proceedings of CHI 2019* (Glasgow, Scotland Uk). Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300376

[45] Ke Li, Ruidong Zhang, Siyuan Chen, Boao Chen, Mose Sakashita, François Guimbretière, and Cheng Zhang. 2024. EyeEcho: Continuous and Low-power Facial Expression Tracking on Glasses. In *Proceedings of the CHI Conference on*

Human Factors in Computing Systems. 1–24.

[46] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. EarIO: A Low-power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 62 (jul 2022), 24 pages. https://doi.org/10.1145/3534621

[47] Longyuan Li, Junchi Yan, Haiyang Wang, and Yaohui Jin. 2020. Anomaly detection of time series with smoothness-inducing sequential variational auto-encoder. *IEEE transactions on neural networks and learning systems* 32, 3 (2020), 1177–1191.

[48] Richard Li, Jason Wu, and Thad Starner. 2019. TongueBoard: An Oral Interface for Subtle Input. In *Proceedings of the 10th Augmented Human International Conference 2019* (Reims, France) *(AH2019)*. Association for Computing Machinery, New York, NY, USA, Article 1, 9 pages. https://doi.org/10.1145/3311823.3311831

[49] Rochelle Lieber. 2009. *Point and manner of articulation of English consonants and vowels*. Cambridge University Press, xii–xii.

[50] Mona Lindau. 1978. Vowel features. *Language* 54, 3 (1978), 541–563.

[51] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. 2021. Wavoice: A Noise-resistant Multi-modal Speech Recognition System Fusing mmWave and Audio Signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems* (Coimbra, Portugal) *(SenSys '21)*. Association for Computing Machinery, New York, NY, USA, 97–110. https://doi.org/10.1145/3485730.3485945

[52] Yu-Ting Liu, Jen-Jee Chen, Yu-Chee Tseng, and Frank Y Li. 2022. An auto-encoder multitask LSTM model for boundary localization. *IEEE Sensors Journal* 22, 11 (2022), 10940–10953.

[53] Andrew J Lotto, Gregory S Hickok, and Lori L Holt. 2009. Reflections on mirror neurons and speech perception. *Trends in cognitive sciences* 13, 3 (2009), 110–114.

[54] Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2023. LLM-Powered Conversational Voice Assistants: Interaction Patterns, Opportunities, Challenges, and Design Guidelines. *arXiv preprint arXiv:2309.13879* (2023).

[55] Hiroyuki Manabe, Akira Hiraiwa, and Toshiaki Sugimura. 2003. Unvoiced speech recognition using EMG-mime speech recognition. In *CHI'03 extended abstracts on Human factors in computing systems*. 794–795.

[56] Scott McGlashan and Tomas Axling. 1996. A speech interface to virtual environments. *Swedish Institute of Computer Science* (1996).

[57] 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.

[58] Meta. 2024. Quest. https://www.meta.com/help/quest/articles/in-vr-experiences/oculus-features/using-voice-commands/

[59] Microsoft. 2024. Cortana. https://learn.microsoft.com/en-us/hololens/hololens-cortana

[60] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. [n. d.]. Large Language Models: A Survey. https://doi.org/10.48550/arXiv.2402.06196 arXiv:2402.06196 [cs]

[61] Nobuyuki Otsu et al. 1975. A threshold selection method from gray-level histograms. *Automatica* 11, 285-296 (1975), 23–27.

[62] Laxmi Pandey and Ahmed Sabbir Arif. 2021. *LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445565

[63] 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*. 1–23.

[64] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 251, 13 pages. https://doi.org/10.1145/3411764.3445430

[65] Laxmi Pandey, Khalad Hasan, and Ahmed Sabbir Arif. 2021. *Acceptability of Speech and Silent Speech Input Methods in Private and Public*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445430

[66] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779* (2019).

[67] Yuru Pei, Tae-Kyun Kim, and Hongbin Zha. 2013. Unsupervised random forest manifold alignment for lipreading. In *Proceedings of the IEEE International Conference on Computer Vision*. 129–136.

[68] Philips. 2024. Philips Samrt Bulbs. https://t.ly/Q4JED

[69] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

[70] L.R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 2 (1989), 257–286. https://doi.org/10.1109/5.18626

[71] Lawrence Rabiner and Biing-Hwang Juang. 1993. *Fundamentals of speech recognition*. Prentice-Hall, Inc., USA.

[72] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[73] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[74] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[75] Jun Rekimoto and Yu Nishimura. 2021. Derma: Silent Speech Interaction Using Transcutaneous Motion Sensing. In *Augmented Humans Conference 2021* (Rovaniemi, Finland) *(AHs'21)*. Association for Computing Machinery, New York, NY, USA, 91–100. https://doi.org/10.1145/3458709.3458941

[76] 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, 749–752.

[77] Rick M Roark. 2006. Frequency and voice: perspectives in the time domain. *Journal of Voice* 20, 3 (2006), 325–354.

[78] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. 2022. Sensing with earables: A systematic literature review and taxonomy of phenomena. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 6, 3 (2022), 1–57.

[79] Stuart Rosen. 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 336, 1278 (1992), 367–373.

[80] Sherry Ruan, Jacob O Wobbrock, Kenny Liou, Andrew Ng, and James A Landay. 2018. Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–23.

[81] Himanshu Sahni, Abdelkareem Bedri, Gabriel Reyes, Pavleen Thukral, Zehua Guo, Thad Starner, and Maysam Ghovanloo. 2014. The Tongue and Ear Interface: A Wearable System for Silent Speech Recognition. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers* (Seattle, Washington) *(ISWC '14)*. Association for Computing Machinery, New York, NY, USA, 47–54. https://doi.org/10.1145/2634317.2634322

[82] Khairul Khaizi Mohd Shariff, Auni Nadiah Yusni, Mohd Adli Md Ali, Megat Syahirul Amin Megat Ali, Megat Zuhairy Megat Tajuddin, and MAA Younis. 2022. Cw radar based silent speech interface using CNN. In *2022 IEEE Symposium on Wireless Technology & Applications (ISWTA)*. IEEE, 76–81.

[83] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. MuteIt: Jaw Motion Based Unvoiced Command Recognition Using Earable. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 140 (sep 2022), 26 pages. https://doi.org/10.1145/3550281

[84] Tanmay Srivastava, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2024. Jawthenticate: Microphone-free Speech-based Authentication using Jaw Motion and Facial Vibrations. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems* (<conf-loc>, <city>Istanbul</city>, <country>Turkiye</country>, </conf-loc>) *(SenSys '23)*. Association for Computing Machinery, New York, NY, USA, 209–222. https://doi.org/10.1145/3625687.3625813

[85] Tanmay Srivastava, R. Michael Winters, Yu-Te Wang, Thomas M. Gable, Teresa LaScala, and Ivan J. Tashev. 2024. *Whispering Wearables*: Multimodal Approach to Silent Speech Recognition with Head-Worn Devices. In *Proceedings of the INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24)* (San Jose, Costa Rica). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3678957.3685720 This work was done when the author was interning at Microsoft Research, Redmond..

[86] 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*. 581–593.

[87] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. [n. d.]. Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision. https://doi.org/10.48550/arXiv.2305.03047 arXiv:2305.03047 [cs]

[88] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. 4214–4217. https://doi.org/10.1109/ICASSP.2010.5495701

[89] J. Tan, C. Nguyen, and X. Wang. 2017. SilentTalk: Lip reading through ultrasonic sensing on mobile phones. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*. 1–9. https://doi.org/10.1109/INFOCOM.2017.8057099

[90] Asterios Toutios and Shrikanth S Narayanan. 2016. Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research. *APSIPA Transactions on Signal and Information Processing* 5 (2016), e6.

[91] Toyota. 2024. Toyota Voice Control. https://toyota-en-us.visteoninfotainment.com/how-to-voice-recognition

[92] Punitha Vancha, Harshitha Nagarajan, Vishnu Sai Inakollu, Deepa Gupta, and Susmitha Vekkot. 2022. Word-level speech dataset creation for sourashtra and recognition system using kaldi. In *2022 IEEE 19th India Council International Conference (INDICON)*. IEEE, 1–6.

[93] Pierre-Hugo Vial, Paul Magron, Thomas Oberlin, and Cédric Févotte. 2021. Phase retrieval with Bregman divergences and application to audio signal recovery. *IEEE Journal of Selected Topics in Signal Processing* 15, 1 (2021), 51–64.

[94] M Vimalkumar, Sujeet Kumar Sharma, Jang Bahadur Singh, and Yogesh K Dwivedi. 2021. 'Okay google, what about my privacy?': User's privacy perceptions and acceptance of voice based digital assistants. *Computers in Human Behavior* 120 (2021), 106763.

[95] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I. Hong, Carmel Majidi, and Swarun Kumar. 2019. RFID Tattoo: A Wireless Platform for Speech Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 155 (Dec. 2019), 24 pages. https://doi.org/10.1145/3369812

[96] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I. Hong, Carmel Majidi, and Swarun Kumar. 2020. RFID Tattoo: A Wireless Platform for Speech Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 155 (sep 2020), 24 pages. https://doi.org/10.1145/3369812

[97] 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*. 1–15.

[98] Jinghan Wu, Yakun Zhang, Liang Xie, Ye Yan, Xu Zhang, Shuang Liu, Xingwei An, Erwei Yin, and Dong Ming. 2022. A novel silent speech recognition approach based on parallel inception convolutional neural network and Mel frequency spectral coefficient. *Frontiers in Neurorobotics* 16 (2022), 971446.

[99] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. LIMU-BERT: Unleashing the Potential of Unlabeled Data for IMU Sensing Applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems* (Coimbra, Portugal) *(SenSys '21)*. Association for Computing Machinery, New York, NY, USA, 220–233. https://doi.org/10.1145/3485730.3485937

[100] Lanyu Xu, Arun Iyengar, and Weisong Shi. 2020. CHA: A caching framework for home-based voice assistant systems. In *2020 IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE, 293–306.

[101] Yamaha. 2024. Yamaha Sound Bar. https://usa.yamaha.com/products/audio_visual/sound_bar/ats-2090/index.html

[102] Wai Chee Yau, Sridhar Poosapadi Arjunan, and Dinesh Kant Kumar. 2008. Classification of voiceless speech using facial muscle activity and vision based techniques. In *TENCON 2008-2008 IEEE Region 10 Conference*. IEEE, 1–6.

[103] Jianping Yue. 2006. Spatial visualization by isometric drawing. In *Proceedings of the2006 IJMEINTERTECH Conference, Union, New Jersey,* Vol. 3.

[104] Asri Rizki Yuliani, M Faizal Amri, Endang Suryawati, Ade Ramdan, and Hilman Ferdinandus Pardede. 2021. Speech enhancement using deep learning methods: A review. *Jurnal Elektronika dan Telekomunikasi* 21, 1 (2021), 19–26.

[105] 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.

[106] 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*. 60–65.

[107] Ruidong Zhang, Mingyang Chen, Benjamin Steeper, Yaxuan Li, Zihan Yan, Yizhuo Chen, Songyun Tao, Tuochao Chen, Hyunchul Lim, and Cheng Zhang. 2022. SpeeChin: A Smart Necklace for Silent Speech Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 192 (dec 2022), 23 pages. https://doi.org/10.1145/3494987

[108] 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*. 1–18.

[109] 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.

[110] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536* (2019).