

SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks

Naoki Kimura

The University of Tokyo
Tokyo, Japan

kimura-naoki@g.ecc.u-tokyo.ac.jp

Michinari Kono

The University of Tokyo
Tokyo, Japan
mchkono@acm.org

Jun Rekimoto

The University of Tokyo
Sony Computer Science Laboratories
Tokyo, Japan
rekimoto@acm.org



Figure 1: SottoVoce silent voice system: an ultrasonic echo probe attached under the jaw that reads the internal situation while the user is speaking without actually emitting a voice. By recognizing ultrasound images using deep convolutional neural networks, the user’s voice is resynthesized and can be used to control the existing speech interaction systems such as smart speakers.

ABSTRACT

The availability of digital devices operated by voice is expanding rapidly. However, the applications of voice interfaces are still restricted. For example, speaking in public places becomes an annoyance to the surrounding people, and secret information should not be uttered. Environmental noise may reduce the accuracy of speech recognition. To address these limitations, a system to detect a user’s unvoiced utterance is proposed. From internal information observed by an ultrasonic imaging sensor attached to the underside of the jaw, our proposed system recognizes the utterance contents without the user’s uttering voice. Our proposed deep neural network model is used to obtain acoustic features from a sequence of ultrasound images. We confirmed that audio signals generated by our system can control the existing smart speakers. We also observed that a user can adjust their oral movement to learn and improve the accuracy of their voice recognition.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-5970-2/19/05...\$15.00

<https://doi.org/10.1145/3290605.3300376>

CCS CONCEPTS

• Computing methodologies → Neural networks; • Human-centered computing → Interface design prototyping; Mobile devices; Sound-based input / output; • Social and professional topics → People with disabilities;

KEYWORDS

silent speech, ultrasonic imaging, deep neural networks, human-AI integration

ACM Reference Format:

Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*, May 4–9, 2019, Glasgow, Scotland Uk. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3290605.3300376>

1 INTRODUCTION

Smart devices that are controllable by speech (i.e., speech interaction) are being used in many situations. Smartphones, smart speakers, car-navigation systems, and various home appliances can be controlled by speech and have presented many interactive potentials [47]. Speech interaction does not require visual attention, and it can be used in a dark environment such as a bedroom. Owing to the recent progress of speech-recognition technology and the naturalness of speech synthesis, speech interaction is becoming an irreplaceable process in human-computer interaction. Speech can also

be used while the user of a speech-interaction device is performing other tasks, such as driving, cooking, homeworking, or using a traditional personal computer. For example, while the user is concentrating on a computer-screen and interaction devices such as a keyboard and a mouse, they can still operate other devices with voice interaction.

However, as for the speech interface, two challenges must be overcome. First, using the interface in public places presents limitations. In addition to being an annoyance to the surrounding people, disclosing personal information or secret information by uttering it in public is risky in terms of information security. Next, it cannot be used in a noisy environment, because the accuracy of speech recognition may be declined. These issues are particularly acute when trying to use a speech interface to interact with wearables or mobile computers.

To overcome these challenges, research on “silent-speech recognition” has been conducted [8]. For example, by applying a method known as lip reading, images of the mouth of the speaker or the entire face are captured by a camera, and the content of the utterance is estimated from those images [52]. If the user could simply *mouth* the utterance without actually voicing it, it would be possible to use such a voice interaction in public places. However, with the camera method, it is necessary to install a camera in front of the face, and its form factor renders it unsuitable for wearables or mobile applications. Other approaches of the studies on “non-audible murmur” (NAM) [22, 42] attempts to recognize utterances with a microphone or an accelerometer worn on the skin or throat of the user. In this case, the user speaks with articulated respiratory sound without vocal-fold vibration (namely, whispering). However, to be recognized accurately by the system, a user’s whispering voice tends to be noticeable by other people nearby. Furthermore, some studies attempt to estimate speech by estimating the movement of muscles near the oral cavity by electromyography (EMG) [33, 49]. However, the estimation of free utterances with EMG is still difficult; instead, it is a type of gesture recognition using the movement of the oral cavity. Thus, the number of detectable commands is limited, and the user has to learn new gesture skills instead of using their existing speaking skills.

Instead of using the above-described approaches, we focus on using ultrasonic imaging [11]. Ultrasonic-imaging technology recognizes the internal status in the body by measuring the reflection time of ultrasonic waves radiated into the body. This technology is widely used to grasp the condition of the internal organs for medical purposes. In recent years, small and lightweight systems that can be directly connected to smartphones (e.g., Vscan Extend, General Electronic Company) have appeared. If it were possible to attach a small ultrasonic imaging head around the neck to sense

the situation in the oral cavity and convert it to acoustic information, it would be a useful device for communicating with speech-capable devices without actually speaking aloud; namely, “silent voice interaction” would be possible.

Silent voice interaction by ultrasonic imaging demonstrates two potential advantages over other approaches. First, the ultrasonic-imaging head can be miniaturized, and it can constitute a device with an inconspicuous shape such as a collar. It would be an important feature for designing wearable silent-voice systems. Next, by recognizing the situation in the oral cavity, it would be possible to measure the movement of the tongue, which cannot be observed from the outside. It would thus be possible to reproduce sound more accurately.

Studies have been conducted on silent speech using ultrasound imaging, but many of them are used in combinations with lip or face images; therefore, a camera must be placed in front of the user [26]. This configuration presents a limitation when it is used as a wearable interface device. Recent researchers have challenged to use deep neural networks with ultrasound imaging for silent speech [7, 51]; however, they are not based on convolutional neural networks and are not validated with data retrieved from a mouth movement without speaking. They are only validated with mouth movement when a user is actually emitting a voice. We present a significant step forward using convolutional neural networks and proof-of-concept validations via actual silent speech, interacting with an unchanged smart speaker (Amazon Alexa).

Herein, a silent-voice interaction system-called “SottoVoce” based only on ultrasonic images is described (Figure 1). By combining two types of deep neural networks, this system can be trained to generate voice signals from a sequence of images captured from an ultrasonic-imaging device. Our contributions can be summarized as the following two topics:

- A two-level model of deep convolutional neural networks to convert ultrasonic images to actual sounds is proposed.
- As a proof of concept, a silent-voice system was developed, and it was shown that the system could control a voice-controlled device (in this case, Amazon Echo) without modifications.

2 RELATED WORK

Silent Speech

Silent-speech interfaces have been studied using various technologies and methods [8, 31]. Lip reading [52] or facial images [12] can be used to estimate speech uttered by a subject without using audio information.

SilentVoice [17] is an *ingressive speech* approach that captures extremely soft speech. Electromagnetic articulography (EMA) has been used to develop brain-computer interfaces [5] and other interactive applications [53]. Magnets

can be attached to the subject to detect silent speech [14, 19, 23]. Electroencephalogram (EEG) [44] and Electromyography (EMG) [37, 49] are also typical methods for silent-speech recognition. Particularly, EMG has been applied for interactive purposes and applications concerning human-computer interaction (HCI) [38]; for example, controlling a web browser [32]. A recent example proposed by Kapur et al. [33] used multiple electrodes to sense neuromuscular signals of a subject in an internal speech. In addition to the methods introduced above, other human-facial electrical potentials have been combined and measured [22, 48].

Ultrasound imaging has also been used for silent-speech recognition in the field of speech processing [9, 27]. In one study, which focused on singing, sung vowels could be synthesized based on the ultrasound and video of the lips [30]. In a similar study, a silent-speech interface using ultrasound and optical images of the tongue and lips was developed [26, 28]. In another study, a mapping technique for automatically generating animations of the tongue movement from raw ultrasound images was created [13].

Other approaches combine the techniques described above with deep neural networks, such as BCI applications [4], myoelectric signals [10], and lip reading with long short-term memory (LSTM) [52]. Combining deep neural networks and ultrasound imaging for silent-speech interfaces has also been considered. For example, ultrasound imaging was used to capture the tongue movement with such a combination with deep neural networks [7]. Moreover, the fundamental frequency (F0) curve, which had been considered unpredictable, was estimated using a deep neural network [21]. In addition, it has been suggested that a global and visuo-acoustic modeling approach called “Eigentongues” performs better than tongue-contour modeling when using neural networks [25]. A new benchmark for silent-speech research based on deep neural networks has also been proposed [31]. These approaches are still at the basic proof-of-concept level, and none have been evaluated in terms of controlling the existing speech-interaction appliances such as smart speakers.

Following the prior approaches described above, we aim to develop a system that allows silent speech to interact with voice-controlled devices and interactive purposes in a HCI context. Furthermore, to improve the performance of the developed system, we aim for the interaction between human and artificial neural networks for performance improvement.

Non-Auditory Inputs and Interaction

EarFieldSensing [39] is a gesture-recognition technology based on electric-field sensing. CanalSense [3] senses changes in air pressure in the ear canals that occur when the face is moved. Tongue-in-Cheek [18] senses the movement of the tongue using the X-band Doppler radar for facial-gesture

recognition. Other methods that use EMG [55] or combinations of brain and muscle signal sensing [43] are also notable.

The techniques above have been utilized for arm-gesture recognition; however, an approach using ultrasound imaging for that purpose has been demonstrated, i.e., EchoFlex [40]. It is an interaction sensor that recognizes movements of the forearm muscle using ultrasound imaging. The results of that study indicated that the sensor performed well and can potentially be supplanted to other prior approaches. They also inspired the authors to consider using ultrasound imaging for silent-voice interaction using the inner part of the mouth.

Interacting with Smart Devices

Owing to the development of mobile and smart devices, we can now interact with them frequently through various methods [35]. We may use our voice as an input method [47], and studies to overcome issues concerning such voice-controlled user interfaces have been presented [41]. One typical interface (a personal “agent” hereafter) is Amazon Echo, typically known as “Alexa,” with which people can communicate and chat. The effect of this agent on people has been researched extensively, and the results show that it is an effective agent for satisfying or influencing our lives [36, 45, 50].

The role of artificial intelligence (AI) has become more important for tasks other than personal agents like Alexa, and we now interact and collaborate with AIs; for example, when sending text messages [24] and designing objects [16]. The hybrid existence and interaction of humans and AI is an impressive topic, which can overcome social issues and improve the quality of our lives. Glove Talk II [15], which was presented in 1995, is a gesture-to-speech system that translates hand gestures into 10 control parameters of a speech synthesizer using neural networks. However, to use the system, the user requires a long-term training of approximately 100 hours. In the present study, referring to this work, we apply the recently improved technologies of deep-neural networks to create the hybrid interaction of humans and AI, where the user (a human) learns to adapt and utilize the AI system to achieve a better interaction.

3 SYSTEM ARCHITECTURE OF SOTTOVOCE

The architecture of the proposed system for generating sound from ultrasonic images is shown in Figure 2. In general, the goal of the system is to transfer certain sequence representations (in this case, ultrasonic images) into other sequence representations (in this case, speech). This goal is similar to that of text-to-speech systems [54], voice-transfer systems [2], and lip reading or face-to-voice systems [12]. Inspired by these systems, the proposed system uses two neural networks.

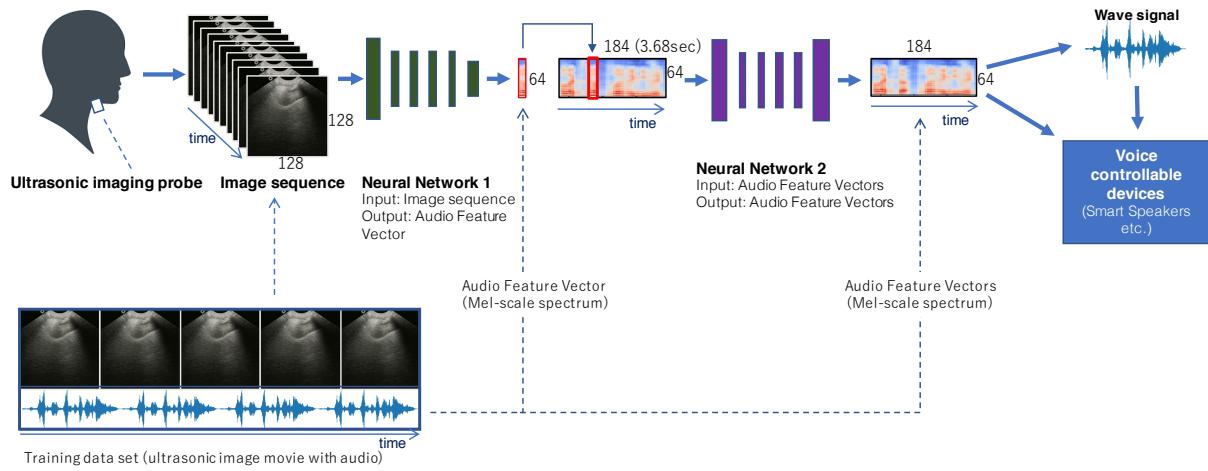


Figure 2: SottoVoce system overview



Figure 3: Ultrasonic imaging probes

The first neural network ('Network 1' in Figure 2) transfers a time series of ultrasonic images to a sound-representation vector. We used a 64-dimensional Mel-scale spectrum with the frequency range of 300 Hz to 8,000 Hz, sampled every 20 ms, as a sound-representation vector. Subsequently, the translated sound representations constitute a series of sound-representation vectors (i.e., a spectrogram). This sound spectrogram can be converted to an audio signal. In addition, to refine the quality of those vectors, they are also transferred by the next neural network ("Network 2" in Figure 2), which generates a series of sound-representation vectors of the same length as that of the input sound-representation vectors. Finally, the output vectors are converted to an actual audio signal.

These two networks are speaker dependent; accordingly, to train them, the system requires a set of ultrasonic-imaging videos captured while the user speaks various speech commands.

Ultrasonic Imaging Device

The CONTEC CMS600P2 Full Digital B-Ultrasound Diagnostic System was used as the ultrasonic-imaging device. A user attaches a 3.5-MHz convex-type ultrasonic imaging probe under the jaw (Figure 3). This system provides a screen output port to be connected to the display monitor. In addition,

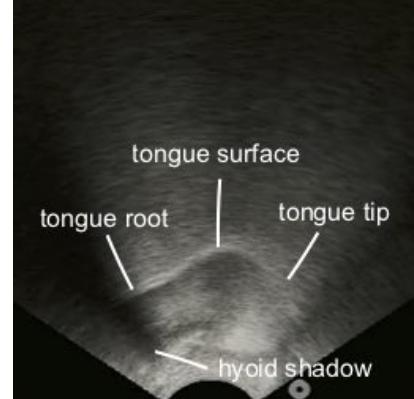


Figure 4: Obtained ultrasonic image from probes attached to the jaw from underneath.

a display-digitizing unit was used for converting the signal sent to the display to an MPEG-4 movie file. Figure 4 shows the obtained ultrasonic image.

We found a delay of ultrasonic images and sound capturing. To compensate for this delay, we examined the corresponding utterance and tongue movement in the video and estimated a delay of 300 ms. We subsequently adjusted to this delay in the training data.

Network 1

Network 1 uses a series of K ultrasonic images (size of 128×128 , monochrome) as the input and generates an n -dimensional sound representation (Mel-scale spectrum) as the output. Currently, K of 13 and n of 64 are used. Because the frame rate of the ultrasonic images is 30 frames per second, the duration of K ultrasonic images is thus 400 ms. This time duration covers the static and motion features of the utterance. Samples of the ultrasonic images are shown in

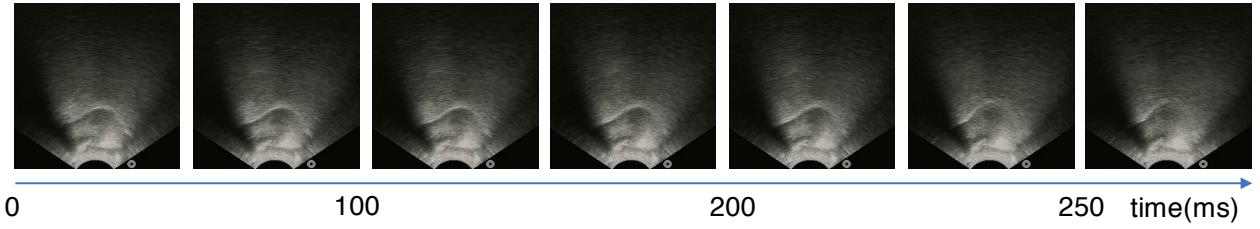


Figure 5: A series of ultrasonic images of the throat of a subject about to pronounce “Alexa.”

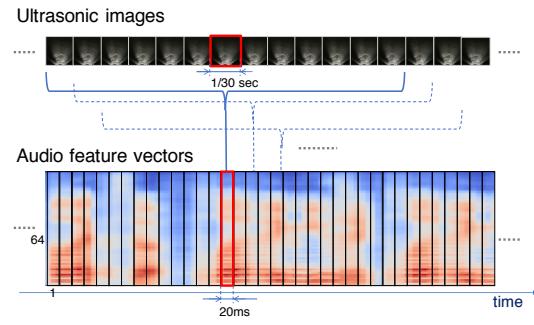


Figure 6: Representation of training data. K-sized ultrasonic images are paired with the corresponding sounds (Mel-scaled sound vectors).

Figure 5. The K-size image sequence is prepared repeatedly such that one Mel-scale spectrum is created every 20 ms (i.e., the hop size is 20 ms) (Figure 6).

A sound-representation vector corresponding to the time position at the center of each ultrasonic-image sequence is extracted from the audio signal data, and Network 1 is trained to generate it.

Network 1 is based on a convolutional neural network (CNN). It comprises four layers: *Conv2D - LeakyReLU - Dropout - BatchNormalization*, followed by six layers: *Flatten - Dense - LeakyReLU - Dropout - Dense - LeakyReLU*. The output size of Network 1 is the same as the length of the sound-representation vector (i.e., 64). Both input images and output vectors are normalized to 0 to one, respectively. The loss function is the mean-squared error, and the optimizer is Adam [34].

Network 2

To improve the sound quality, Network 2 uses a sequence of sound-representation vectors and generates a sequence of sound-representation vectors with the same length as the input. This model comprises a bank of one-dimensional (1-D) convolutional filters (*Conv1D*), with a kernel size from 1 to M ($M = 8$ is currently used), followed by the U-Network [46] with the three layers of *Conv1D - MaxPooling (strides=2) - LeakyReLU - Dropout*, and thrice of *DeConv1D - Concatenate*

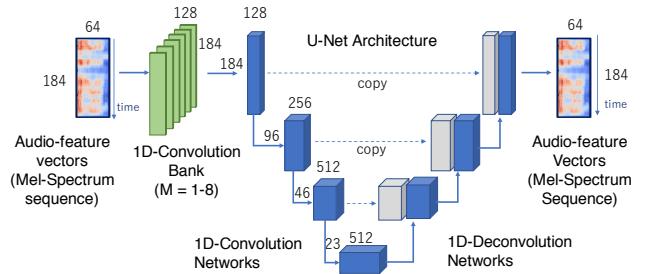


Figure 7: Network 2 improves the quality of generated Mel-scaled spectrum sequences. (Note: for the consistency with the illustration of the neural networks, the time axis of the audio-feature vector is shown as the vertical axis.)

(Figure 7). The 1-D convolutional bank explicitly models the local and contextual information of the input sequence. The following U-Network also improves the quality of the audio sequence with precise localization. Finally, the network generates Mel-scale spectrum vectors.

To train Network 2, Network 1 was used to create Mel-scale spectrum vectors from the images of a training ultrasonic video clip as the input, and the same length Mel-scale spectrum vectors from the audio of the same training video clip as the output. Similarly, in the case of Network 1, the mean-squared error was used as a loss function, and Adam was used as an optimizer.

For simplicity, the time durations of the input and output were fixed to the same value (currently, 3.68 s is used). This duration encompasses many typical speech commands.

Sound generation

Following the neural-networks processing, a sequence of Mel-scale-spectrum sound-representation vectors is converted to an audio signal using the Griffin Lim algorithm [20]. This conversion is possible from the output of Network 1 or the output of Network 2.

For testing, the generated audio signals are transmitted from an audio speaker, and they can be used to control nearby sound-controlled devices such as a smart speaker. We are also considering taking the audio-waveform signal directly

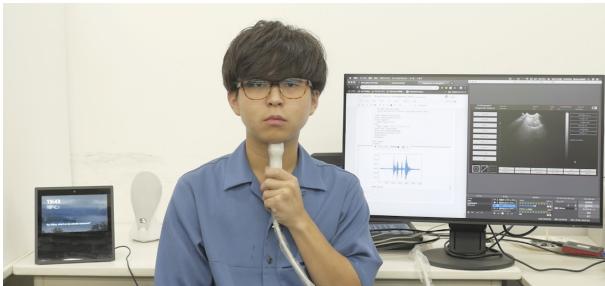


Figure 8: Apparatus used for training and evaluating.

as the audio input information of the speech-controllable device without actually reproducing it as a sound wave.

Training

As for preparing the training data, two collaborators, (28-year old male and 24-year old male) were attached with an ultrasonic imaging probe under their jaws, and were instructed to utter various speech commands. Approximately 500 speech commands were collected from each collaborator (Table ??). For each command, as well as the voice utterance, a video of the ultrasonic images was recorded. The training session was approved by the research ethics committee of the author's institution.

The recorded video was used to train Network 1. The ultrasonic images were rescaled to 128×128 and used as inputs. The corresponding utterance voice was converted to a Mel-scale spectrum and used as outputs.

The number of test sets for Network 2 was the same as the number of trained video files (approximately 500). To increase the number of test sets, data augmentation by applying Gaussian noise to the input Mel-scale spectrum vectors was used.

As our model is speaker dependent, both Network 1 and Network 2 are trained for each speaker. Network 1 is trained first and subsequently used to create the dataset for training Network 2.

Implementation Details

The above-described network models were implemented based on the Keras [6] deep-learning platform with Tensorflow [1] as the backend, and an NVIDIA GeForce 1080ti as the GPU board. Training Network 1 with 500 speech commands (which creates 35,000 training data pairs for Network 1) required approximately 4 h. Training Network 2 required less than an hour.

As the ultrasonic-imaging device cannot be connected directly to the Ubuntu machine that runs the neural networks, a simple server-client program was developed. It connects the computer that controls the ultrasonic imaging device to

the computer that operates the neural networks. The generated audio signals are sent back to the computer with the ultrasonic-imaging device.

To process ultrasonic images at/with duration/intervals of 3.68 s, 2.36 s is required for the neural networks to process the images. The total processing time (including video processing, neural networks processing, and conversion of the Mel-scale spectrum to an audio wave) was 2.61 s.

4 RESULTS

The results of converting the ultrasonic images to sound are shown in Figure 9. In the figure, the top-row graphs show the sound-representation vectors (a Mel-scale spectrogram), and the bottom-row graphs are the corresponding waveforms. The graph labeled Net1 is the result of Network 1, that labeled Net2 is the result of Network 1 + Network 2, and that labeled "original" is the wave data encoded as a Mel-scale spectrogram and decoded back to the waveform. Thus, the last one is regarded as the ground-truth of the training. Although the difference between the outputs of Network 1 and Network 2 was unclear, we observed that the sound generated by Network 2 was better than that generated by Network 1 (Examples of the output audio signals are given in the supplemental video.)

It is noteworthy that both Network 1 and Network 2 generate natural intonation, which is typically considered to be caused by vocal-fold vibration, not by the internal situation. This result might suggest that the neural networks learn the context of the speech.

The generated sounds emitted from the computer's speaker were subsequently tested with an existing (unchanged) smart speaker (Amazon Echo and Amazon Echo Show), and this test confirmed that the generated sounds can control smart speakers. The speech commands used for training and testing were typical Amazon Alexa commands. For this test, the participants spoke the following four commands, five times each (20 utterances in total): "Alexa, play music," "Alexa, what's the weather like," "Alexa, what time is it," and "Alexa, play jazz."

Table 1 lists the recognition success ratio of Network 1, Network 1 + Network 2, and the original (Mel-scale encoded and decoded) as a ground truth. We confirmed that the combination of Network 1 and Network 2 improves the recognition rate. We also noticed that the trigger word ("Alexa") is always regenerated clearly. This may because that word is simply the most pronounced word in the training set.

We also performed the word error rate (WER) measure test using Google's cloud speech-to-text engine [29] and the same environment used for the recognition measurement of smart speakers. The WER was 20.61%, 41.03%, and 33.56% for GT, Network 1, and Network 2, respectively (mean of a total of 40 speech commands from two users from Table 1). We believe

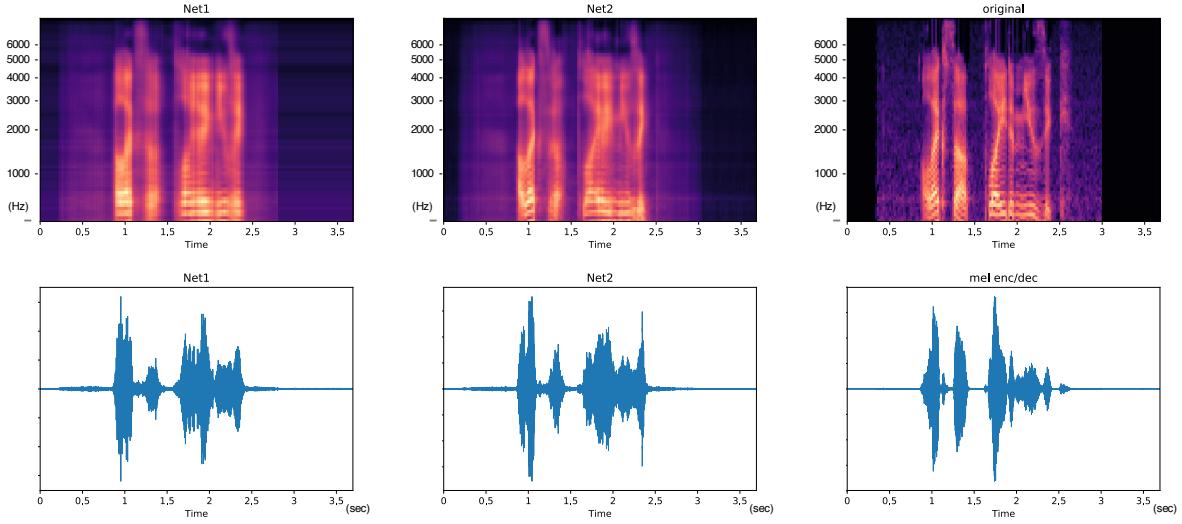


Figure 9: Results of the training (shown as waves and Mel-scale spectrogram): Net1: Network 1 results; Net2: Network 2 results; original: original voice encoded by Mel-scale spectrogram and decoded to an audio signal. This is the “ground-truth” of this training.

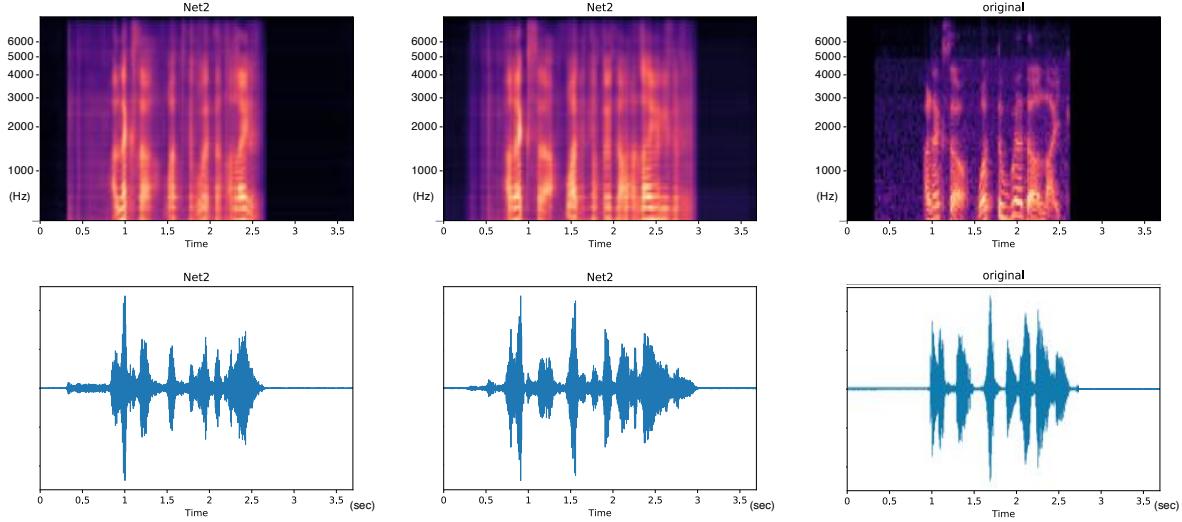


Figure 10: Comparison of Generated Sounds. Left: Network 2 results from ultrasonic images while a user is emitting a voice; Middle: Network 2 results from ultrasonic images without emitting a voice; Right: the original (Mel-scale spectrogram encoded/decoded) voice

that this will also serve as evidence for the effectiveness of Network 2. Through these studies, we found that commands such as “what’s the weather like?” had a high recognition rate for all conditions, while commands with shorter terms such as “play jazz” performed much worse. This may suggest that the longer commands are easier for Network 2 to obtain the contexts.

	User A	User B	ave.
Network 1	60.0%	25.0%	42.5%
Network 1 + Network 2	65.0%	65.0%	65.0%
GT	90.0%	90.0%	90.0%

Table 1: Speech-recognition success ratio in tests with an unchanged existing smart speaker (“GT” means Mel-scale encoded/decoded from the original voice data, regarded as the ground truth of the training).

5 END-TO-END EVALUATION AND OBSERVATIONS

The real end-to-end silent voice to audio conversion was examined. In this case, a user is asked to mouth a speech command without actually emitting a sound, and the oral cavity movement is recorded by an ultrasonic imaging probe. The obtained image sequence is subsequently translated to a voice by the proposed system.

We asked the participants to speak as silently as possible (not to vibrate their vocal cords) and to try to speak as similar as possible when they speak with sound. However, we did not ask them to hold their breaths. Consequently, a case occurred where a small leaked sound was audible. To clarify this situation, we measured the level of the emitted sound level from the participant by following the evaluation method of SilentVoice [17]. We used a noise meter that has the same specifications (min range 30 dB, 1.5 dB error) placed 30 cm away, in a room with a background noise level of 31.0 dB(A). The mean of the peak sound level of 20 measurements (typical Alexa speech commands) was 37.14dB(A), which is lower than that of soft whispering.

Figure 10 shows the comparison of the generated sound when a user is emitting a voice and when a user is not emitting a voice (please refer to the supplemental video for the actual generated sounds).

Initially, the result was unsatisfactory. We first expected that the movement in the oral cavity without emitting a voice (Figure 10 (middle)) is the same as that when the user actually emits a voice (Figure 10 (left)); however, a subtle difference was found between them, and the sound quality generated by the image-without-voice was not as good as that generated by the image-with-voice.

However, the following interesting phenomenon was observed. As the user can also listen to the generated sound generated by the image-without-voice, the user attempted to change the mouth movement to obtain a slightly better result. After several trials, the quality of generated sound improved. In this case, it is considered that the users improved their own skills in silent voicing.

6 DISCUSSIONS

Incremental Voice Generation

In our current design, the obtained ultrasonic-image sequences were converted to a voice at the granularity of the speech command (approximately 3.6 s). This is because Network 2 uses a fixed-length voice-representation sequence. However, based on the observation of the users' practice, it should be better to generate sounds incrementally, such that the user can have a tighter feedback loop for learning oral movements for generating a better voice.



Figure 11: Future Image: possible configuration of an ultrasonic probe and an open earphone

Continuous Ultrasonic-Wave Emission into the Body

The effect on human organs when ultrasonic waves are emitted continuously into the body is unknown. However, we may be able to combine a simple triggering mechanism to start and stop the emission of ultrasonic waves. For example, the combination of an accelerometer and a microphone in the device can detect jaw movements for starting the (silent) voice command without actually emitting a voice.

Cure for Vocal Cord Disabilities

We also expect people with damaged vocal cords to use our research. As described in the Human-AI integration section, people will be able to learn how to correctly control their mouth and tongue to generate sound, even though their vocal cords do not work.

Combining with Other Modalities

Finally, it should be mentioned that this research is not intended to exclude other modalities. Combining the information of EMG, accelerometers, and NAM microphones may improve the quality of speech recognition. Investigating the combination of these modalities is subject of future research.

Human-AI integration

The above-described observation suggests that an interesting relationship exists between humans and AI. Rather than considering AI as an autonomous or separated entity, we may be able to regard AI as a part of humans. Hence, even when the initial performance of the (artificial) neural networks is not perfect, a user can gradually learn and improve their performance.

This is similar to how people learn fundamental skills. When we learn utterance, the coordination of the motor cortex that drives the oral cavity, tongue, and vocal folds, and the auditory cortex forms a tight loop to obtain better

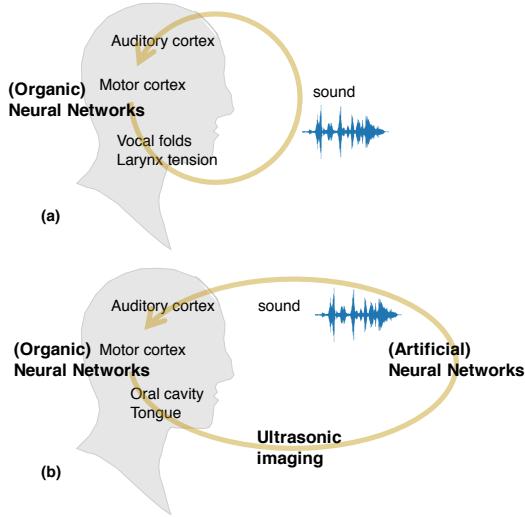


Figure 12: Human–AI integration: artificial neural networks and organic neural networks form a feedback loop for obtaining better results.

speech performance (Figure 12 (a)). By extending this loop, organic neural networks (e.g., our brain) and artificial neural networks may also form a tight feedback loop (Figure 12 (b)). We name this formation, “human–AI integration” rather than human–AI interaction.

In this regard, we refer to the research of Glove Talk II from 1995 a pioneering work on human–AI integration [15]. In this work, the user learned to control a voice synthesizer with hand gestures. Three simple neural networks were used, and the user (who was a pianist) required more than 100 h to generate an audible voice. A combination of better neural networks and a learner could reduce this learning time.

7 CONCLUSION

A method of silent-voice interaction with ultrasonic imaging was proposed. Two neural networks were used in sequence to convert a mouthed “utterance” of a user without a voice to a sound (voice), and could be used to operate the existing voice-controllable devices such as smart speakers.

Following this result, we envision that the future form-factor for the wearable computer would be a combination of an attachable ultrasonic imaging probe to the underside of the jaw, with a bone conductive earphone or an open-air earphone (Figure 11). With this configuration, a user can always invoke a voice-controllable assistant without emitting a voice and obtain responses.

REFERENCES

- [1] Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. <https://www.tensorflow.org/> Software available from tensorflow.org.
- [2] Dabi Ahn. 2017. Voice Conversion with Non-Parallel Data. <https://github.com/andabi/deep-voice-conversion>.
- [3] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2017. CanalSense: Face-Related Movement Recognition System Based on Sensing Air Pressure in Ear Canals. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST ’17)*. ACM, New York, NY, USA, 679–689. <https://doi.org/10.1145/3126594.3126649>.
- [4] F Bocquelet, Thomas Hueber, Laurent Girin, Pierre Badin, and B Yvert. 2014. Robust articulatory speech synthesis using deep neural networks for BCI applications. , 2288–2292 pages.
- [5] Florent Bocquelet, Thomas Hueber, Laurent Girin, Christophe Savariaux, and Blaise Yvert. 2016. Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces. *PLOS Computational Biology* 12, 11 (11 2016), 1–28. <https://doi.org/10.1371/journal.pcbi.1005119>
- [6] François Chollet et al. 2015. Keras. <https://keras.io>.
- [7] Tamás Gábor Csápi, Tamás Grósz, Gábor Gosztolya, László Tóth, and Alexandra Markó. 2017. DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface. In *INTERSPEECH*.
- [8] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg. 2010. Silent Speech Interfaces. *Speech Commun.* 52, 4 (April 2010), 270–287. <https://doi.org/10.1016/j.specom.2009.08.002>
- [9] B. Denby and M. Stone. 2004. Speech synthesis from real time ultrasound images of the tongue. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. I–685. <https://doi.org/10.1109/ICASSP.2004.1326078>
- [10] L. Diener, M. Janke, and T. Schultz. 2015. Direct conversion from facial myoelectric signals to speech using Deep Neural Networks. In *2015 International Joint Conference on Neural Networks (IJCNN)*. 1–7. <https://doi.org/10.1109/IJCNN.2015.7280404>
- [11] Roman Gr. Maev (ed.). 2013. *Advances in Acoustic Microscopy and High Resolution Imaging: From Principles to Applications*. Wiley-VCH.
- [12] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. 2017. Improved Speech Reconstruction from Silent Video. *ICCV 2017 Workshop on Computer Vision for Audio-Visual Media* (2017).
- [13] Diandra Fabre, Thomas Hueber, Laurent Girin, Xavier Alameda-Pineda, and Pierre Badin. 2017. Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract. *Speech Communication* 93 (2017), 63 – 75. <https://doi.org/10.1016/j.specom.2017.08.002>
- [14] M.J. Fagan, S.R. Ell, J.M. Gilbert, E. Sarrazin, and P.M. Chapman. 2008. Development of a (silent) speech recognition system for patients following laryngectomy. *Medical Engineering & Physics* 30, 4 (2008), 419 – 425. <https://doi.org/10.1016/j.medengphy.2007.05.003>
- [15] Sidney Fels and Geoffrey Hinton. 1995. Glove-TalkII: An Adaptive Gesture-to-formant Interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’95)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 456–463. <https://doi.org/10.1145/223904.223966>
- [16] Rebecca Fiebrink. 2017. *Machine Learning as Meta-Instrument: Human-Machine Partnerships Shaping Expressive Instrumental Creation*. Springer Singapore, Singapore, 137–151. https://doi.org/10.1007/978-981-10-2951-6_10
- [17] Masaaki Fukumoto. 2018. SilentVoice: Unnoticeable Voice Input by Ingressive Speech. In *Proceedings of the 31st Annual ACM Symposium*

- on User Interface Software and Technology (UIST '18)*. ACM, New York, NY, USA, 237–246. <https://doi.org/10.1145/3242587.3242603>
- [18] Mayank Goel, Chen Zhao, Ruth Vinisha, and Shwetak N. Patel. 2015. Tongue-in-Cheek: Using Wireless Signals to Enable Non-Intrusive and Flexible Facial Gestures Detection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 255–258. <https://doi.org/10.1145/2702123.2702591>
- [19] Jose A. Gonzalez, Lam A. Cheah, James M. Gilbert, Jie Bai, Stephen R. Ell, Phil D. Green, and Roger K. Moore. 2016. A Silent Speech System Based on Permanent Magnet Articulography and Direct Synthesis. *Comput. Speech Lang.* 39, C (Sept. 2016), 67–87. <https://doi.org/10.1016/j.csl.2016.02.002>
- [20] D. Griffin and Jae Lim. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32, 2 (April 1984), 236–243. <https://doi.org/10.1109/TASSP.1984.1164317>
- [21] Tamás Grósz, Gábor Gosztolya, László Tóth, Tamás Csapó, and Alexandra Markó. 2018. F0 Estimation for DNN-Based Ultrasound Silent Speech Interfaces.
- [22] Tatsuya Hirahara, Makoto Otani, Shota Shimizu, Tomoki Toda, Keigo Nakamura, Yoshitaka Nakajima, and Kiyohiro Shikano. 2010. Silent-speech enhancement using body-conducted vocal-tract resonance signals. *Speech Communication* 52, 4 (2010), 301 – 313. <https://doi.org/10.1016/j.specom.2009.12.001> Silent Speech Interfaces.
- [23] 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 Commun.* 55, 1 (Jan. 2013), 22–32. <https://doi.org/10.1016/j.specom.2012.02.001>
- [24] Jess Hohenstein and Malte Jung. 2018. AI-Supported Messaging: An Investigation of Human-Human Text Conversation with AI Support. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. ACM, New York, NY, USA, Article LBW089, 6 pages. <https://doi.org/10.1145/3170427.3188487>
- [25] T. Hueber, G. Aversano, G. Cholle, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone. 2007. Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Vol. 1. I-1245–I-1248. <https://doi.org/10.1109/ICASSP.2007.366140>
- [26] 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 Commun.* 52, 4 (April 2010), 288–300. <https://doi.org/10.1016/j.specom.2009.11.004>
- [27] Thomas Hueber, Elie-Laurent Benaroya, Bruce Denby, and Gérard Chollet. 2011. Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface. In *INTER-SPEECH*.
- [28] Thomas Hueber, Gerard Chollet, Bruce Denby, and M Stone. 2008. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. (01 2008).
- [29] Google Inc. [n. d.]. Clund Speech-to-Text. <https://cloud.google.com/speech-to-text/>.
- [30] Aurore Jaumard-Hakoun, Kele Xu, Clémence Leboulenger, Pierre Roussel-Ragot, and Bruce Denby. 2016. An Articulatory-Based Singing Voice Synthesis Using Tongue and Lips Imaging. In *ISCA Interspeech 2016*, Vol. 2016. San Francisco, United States, 1467 – 1471. <https://doi.org/10.21437/Interspeech.2016-385>
- [31] Yan Ji, Licheng Liu, Hongcui Wang, Zhilei Liu, Zhibin Niu, and Bruce Denby. 2018. Updating the Silent Speech Challenge Benchmark with Deep Learning. *Speech Commun.* 98, C (April 2018), 42–50. <https://doi.org/10.1016/j.specom.2018.02.002>
- [32] Chuck Jorgensen and Kim Binsted. 2005. Web Browser Control Using EMG Based Sub Vocal Speech Recognition. In *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences - Volume 09 (HICSS '05)*. IEEE Computer Society, Washington, DC, USA, 294.3–. <https://doi.org/10.1109/HICSS.2005.683>
- [33] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. AlterEgo: A Personalized Wearable Silent Speech Interface. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 43–53. <https://doi.org/10.1145/3172944.3172977>
- [34] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). arXiv:1412.6980 <http://arxiv.org/abs/1412.6980>
- [35] Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. 2017. The Rise of Bots: A Survey of Conversational Interfaces, Patterns, and Paradigms. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. ACM, New York, NY, USA, 555–565. <https://doi.org/10.1145/3064663.3064672>
- [36] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 881–894. <https://doi.org/10.1145/3196709.3196784>
- [37] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel. 2005. Session independent non-audible speech recognition using surface electromyography. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005*. 331–336. <https://doi.org/10.1109/ASRU.2005.1566521>
- [38] 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 (CHI EA '03)*. ACM, New York, NY, USA, 794–795. <https://doi.org/10.1145/765891.765996>
- [39] Denys J. C. Matthies, Bernhard A. Strecker, and Bodo Urban. 2017. EarFieldSensing: A Novel In-Ear Electric Field Sensing to Enrich Wearable Gesture Input Through Facial Expressions. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1911–1922. <https://doi.org/10.1145/3025453.3025692>
- [40] Jess McIntosh, Asier Marzo, Mike Fraser, and Carol Phillips. 2017. EchoFlex: Hand Gesture Recognition Using Ultrasound Imaging. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1923–1934. <https://doi.org/10.1145/3025453.3025807>
- [41] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 6, 7 pages. <https://doi.org/10.1145/3173574.3173580>
- [42] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, Vol. 5. V–708. <https://doi.org/10.1109/ICASSP.2003.1200069>
- [43] Phuc Nguyen, Nam Bui, Anh Nguyen, Hoang Truong, Abhijit Suresh, Matt Whitlock, Duy Pham, Thang Dinh, and Tam Vu. 2018. TYTH-Typing On Your Teeth: Tongue-Teeth Localization for Human-Computer Interface. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '18)*. ACM, New York, NY, USA, 269–282. <https://doi.org/10.1145/3210240.3210322>

- [44] Anne Porbadnigk, Marek Wester, Jan Calliess, and Tanja Schultz. 2009. EEG-based Speech Recognition - Impact of Temporal Effects. In *BIOSIGNALS*.
- [45] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa is My New BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 2853–2859. <https://doi.org/10.1145/3027063.3053246>
- [46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR* abs/1505.04597 (2015). arXiv:1505.04597 <http://arxiv.org/abs/1505.04597>
- [47] Alexander I. Rudnicky. 1989. The Design of Voice-driven Interfaces. In *Proceedings of the Workshop on Speech and Natural Language (HLT '89)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 120–124. <https://doi.org/10.3115/100964.100972>
- [48] 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 (ISWC '14)*. ACM, New York, NY, USA, 47–54. <https://doi.org/10.1145/2634317.2634322>
- [49] Tanja Schultz. 2010. ICCHP Keynote: Recognizing Silent and Weak Speech Based on Electromyography. In *Computers Helping People with Special Needs*, Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 595–604.
- [50] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 857–868. <https://doi.org/10.1145/3196709.3196772>
- [51] László Tóth, Gábor Gosztolya, Tamás Grósz, Alexandra Markó, and Tamás Csapó. 2018. Multi-Task Learning of Speech Recognition and Speech Synthesis Parameters for Ultrasound-based Silent Speech Interfaces.
- [52] Michael Wand, Jan Koutník, and Jürgen Schmidhuber. 2016. Lipreading with Long Short-Term Memory. *CoRR* abs/1601.08188 (2016). arXiv:1601.08188 <http://arxiv.org/abs/1601.08188>
- [53] Jun Wang, Ashok Samal, and Jordan Green. 2014. Preliminary Test of a Real-Time, Interactive Silent Speech Interface Based on Electromagnetic Articulograph. (06 2014).
- [54] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model. *CoRR* abs/1703.10135 (2017). arXiv:1703.10135 <http://arxiv.org/abs/1703.10135>
- [55] Qiao Zhang, Shyamnath Gollakota, Ben Taskar, and Raj P.N. Rao. 2014. Non-intrusive Tongue Machine Interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2555–2558. <https://doi.org/10.1145/2556288.2556981>