SilentVoice: Unnoticeable Voice Input by Ingressive Speech

Masaaki FUKUMOTO
Microsoft Research
Beijing, China
fukumoto@microsoft.com

ABSTRACT

SilentVoice is a new voice input interface device that penetrates the speech-based natural user interface (NUI) in daily life. The proposed "ingressive speech" method enables placement of a microphone very close to the front of the mouth without suffering from pop-noise, capturing very soft speech sounds with a good S/N ratio. It realizes ultra-small (less than 39dB(A)) voice leakage, allowing us to use voice input without annoying surrounding people in public and mobile situations as well as offices and homes. By measuring airflow direction, SilentVoice can easily be separated from normal utterances with 98.8% accuracy; no activation words are needed. It can be used for voice-activated systems with a specially trained voice recognizer; evaluation results yield word error rates (WERs) of 1.8% (speaker-dependent condition), and 7.0% (speaker-independent condition) with a limited dictionary of 85 command sentences. A whisper-like natural voice can also be used for real-time voice communication.

Author Keywords

silent voice input; silent speech input; ingressive speech; unnoticeable; awareLESS; SilentVoice; wearable interface; interface devices; wearables; SilentPhone;

CCS Concepts

• **Hardware~Sound-based input / output**
• **Human-centered computing~Sound-based input / output**
• Computing methodologies~Speech recognition

INTRODUCTION

Voice input is one of ideal NUI (natural user interface) [33] for enabling high-speed input without special training [6], and is already widely used especially for specific situations such as interactive voice response (IVR) systems, hands-free operations while driving and package-handling, and clinical record dictations by doctors [28]. Recently, new generations of deep neural network (DNN) and cloud-based voice recognition engines have improved recognition accuracy [5,31,34]; smartphones [3,24] and smart speakers [1] also incorporate voice input interfaces into daily life.

However, we rarely see people using voice input in public spaces and offices[1]. One major reason is that voice leakage annoys surrounding people, and may risk disseminating private information to unintended audiences. These are not technical issues but social issues for which there is no easy fix even if performance of the voice recognition system is greatly improved.

RELATED WORKS

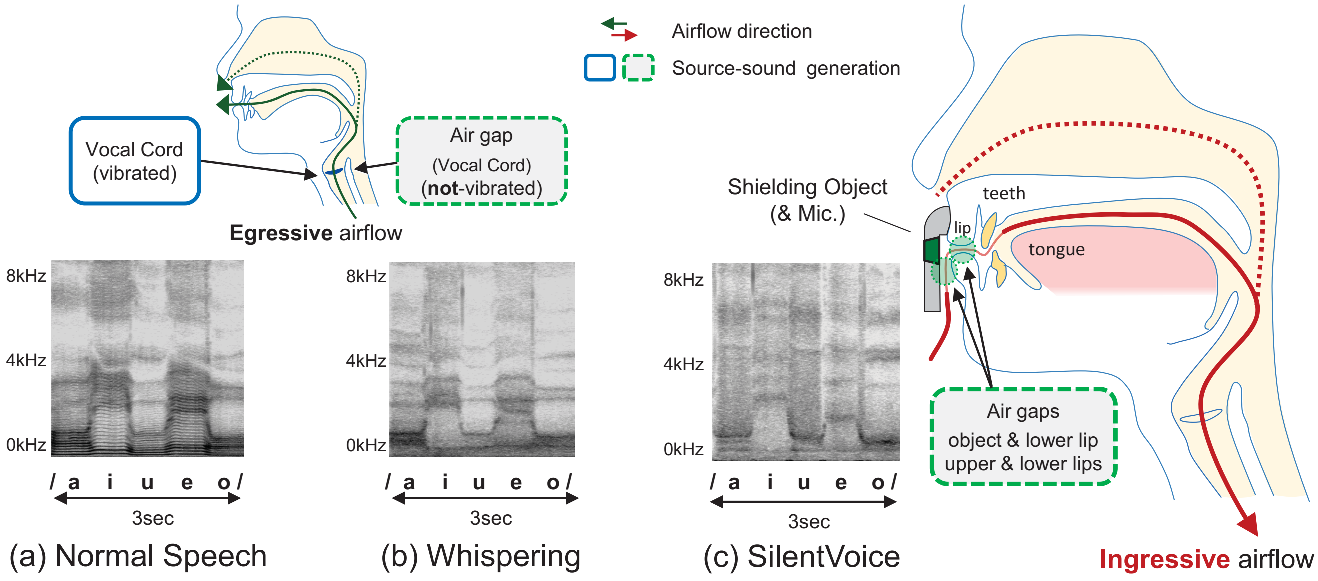
There exist some "silent" voice input systems that can be used without being noticed by surroundings [8]. Soft whisper voices can be detected by using a stethoscopic microphone that contacts the skin behind the ear (called NAM: Non-Audible Mummer) [26], or a throat microphone [17]. Contact microphones can eliminate external noise interference, but it is difficult to reduce spike noise that is often generated in association with normal body movements. Moreover, long-term use of whisper voice might have negative effects on our vocal cords [30]. Another approach is to completely cover the user's mouth by the hood [14], but this makes our voices muffled and looks somewhat strange.

There are some other voice input methods that can work under completely silent conditions; for example, detecting EMG (electromyogram) signals when speaking by placing skin surface electrodes at the face [18,20] or throat [2]; detecting tongue movements by tracking small magnets pasted on the tongue [4]; lip reading by face cam [21]; using ultrasound CT-images of the oral cavity or the larynx [13]; and detecting "imagined" speech by using a multi-channel EEG (electroencephalogram) [19]. These methods can work without generating any sounds, but often require large equipment or can only detect very limited phonemes or words. Therefore, none of them have been widely used as a primary voice input method for normal people.

Voice control is one of the main target applications of "silent" voice systems. However, conventional voice recognizers are basically designed for our normal speech, and cannot recognize other types of speech properly. There exist some special recognizers for whispering [16] or NAM

UIST '18, October 14–17, 2018, Berlin, Germany
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5948-1/18/10…$15.00
https://doi.org/10.1145/3242587.3242603

[1] Especially in Japan, no one uses voice assistant and even cellular phone headsets in public space because they are very unwilling to annoy surrounding individuals. Voice input is also rarely used in their offices and homes due to small and non-private spaces.



(a) Normal Speech (b) Whispering (c) SilentVoice

Figure 1: Structure of utterances with spectrograms (/aiueo/): (a) normal speech, (b) whispering, (c) SilentVoice In SilentVoice, air gaps generate a whisper-like source sound along with ingressive airflow.

voice [26] by training an acoustic model with targeted speech data. Real-time voice communication is another possible application, used in some voice conversion systems such as from whispering to normal voice [19] or from NAM voice to whispering [32]. NAM voice is also used as a speaking aid for total laryngectomees [27].

INGRESSIVE SPEECH

For realizing simple and effective "silent" voice interaction, we propose a novel "ingressive speech"-based voice input method.

Figure 1 shows typical utterance mechanisms with the spectrogram of /aiueo/. In the case of a normal speech (Figure 1(a)), our vocal cords are vibrated by expiratory (or exhaust) air, generating a glottal source sound (about 130Hz of triangular wave in an adult male). Then the source sound resonates at the vocal tract that consists of the larynx, pharynx, vocal cavity, nasal cavity, and paranasal cavity. Finally, it takes on some frequency peaks called formants, which are key elements in recognizing a generated sound as a "human voice".

Figure 1(b) shows the mechanism for whispering. The vocal cords do not vibrate; instead a noise-like source sound is generated by turbulence when expiratory air passes through the narrow gap between the vocal cords (and also passes through the vocal tracts). Like normal speech, the source sound resonates at the vocal tract and gets formants. A whisper has richer high-frequency components and weaker formants compared with a normal voice, but we (humans) can still recognize both sounds as human voices[2].

Consonants are mainly generated by turbulence when expiratory air passes through the vocal cavity (space enclosed by teeth, tongue, and lips), and when expiratory air is blocked (and released) by lips or tongue (e.g. /t/, /p/ sounds). The same methods are used in both normal and whispering voices.

Usually, our utterance is performed while exhaling (breathing-out)[3]. However, similar sound can also be generated by opposite air stream. As an example, some whisper-like sounds are observed when doing same lip, jaw, and tongue movements of "Hello everyone" while **inhaling (breathing-in)**. In this case, air turbulence in the vocal cavity and generated consonants are almost the same in normal voice and whispering, because similar movements of articulation mechanisms are used. On the other hand, the sound source is generated from another kind of air turbulence when inspiratory (or suction) air passes through narrow gaps[4]. These air turbulences are basically the same as that generated at the vocal cords when whispering, therefore, we can recognize an ingressive speech sound as a "whisper-like" voice[5].

This ingressive speech has some advantages (the details will be discussed later).

[2] "Buccal speech" and "pharyngeal speech" are other alaryngeal speech methods [10]; source sound is generated by using the tongue, cheek, and jaw in the vocal cavity. These methods also use expiratory air, and generated phonemes are limited and unclear.

[3] Some language groups have a limited number of phonemes generated by "ingressive speech", such as /ja/ of Swedish [9].

[4] It is effective in making artificial narrow gaps by placing some object (e.g. finger or hand) in front of the mouth, or slightly opened lips because much stronger turbulence is generated with little airflow. See also instruction sheet for subjects in Appendix.

[5] It is not so easy to stably vibrate our vocal cords while inhaling. Cats have special types of vocal cords (called ventricular cords or "false" vocal cords) for generating continuous ingressive "purring" sounds [29].

- No "pop-noise" is generated so that the microphone can be placed very close to the mouth, achieving a high S/N ratio.
- Normal speech and ingressive speech can be separated by simply measuring airflow direction.

By using ingressive speech, ultra-small utterances can be captured with an ultra-closely placed microphone, and we can use voice input not only in quiet rooms but also noisy public spaces without annoying other people.

The following sections describe the details of our ingressive speech-based voice input mechanism (named SilentVoice). The structure and advantages of SilentVoice is described while comparing with normal voice and whispering; leaking sound level and recognition results are also evaluated; possible implementations with application examples are then presented, followed by discussions; the last section concludes the paper.

SILENTVOICE

This section describes the structure of SilentVoice and outlines how it can achieve both low leaking sound and high signal levels with evaluation results.

Structure of Utterance Mechanisms

The structure of SilentVoice is shown in Figure 1(c). A plate or stick-shaped shielding object is placed in front of the mouth with a narrow air gap (e.g. 1mm). Contrary to normal voice and whispering, airflow direction is from outside to inside (lungs). Air turbulence (sound source) is generated by the narrow air gap. By lightly contacting an upper part of the shielding object with the upper lip, the gap shape remains stable even when the mouth moves while speaking. It is also effective in making another air gap with narrowly opened lips while uttering. The generated sound has much richer high-frequency components compared with whispering, and formants are slightly shifted to the high-frequency side, but it still has similar characteristics as whispering. Possible reasons are both methods have similar "turbulence"-based sound sources, and they have the same components for making formants and consonants such as resonance at the vocal tract and articulation mechanisms. In addition, the shielding object should not come in contact with the lower lip or lower jaw during utterance. If touched, it may block movement of the mouth and lower jaw and muffle the voice; touching and sliding sounds also generate unexpected noise.

Advantages of Ingressive Speech

One of the major characteristics of SilentVoice is its use of inspiratory air whereas other methods use expiratory air. In fact, the method shown in Figure 1(c) can generate similar sounds regardless of airflow directions such as inspiratory or expiratory (cf. ingressive whistle, harmonica).

There are a few reasons why we use inspiratory air for SilentVoice.

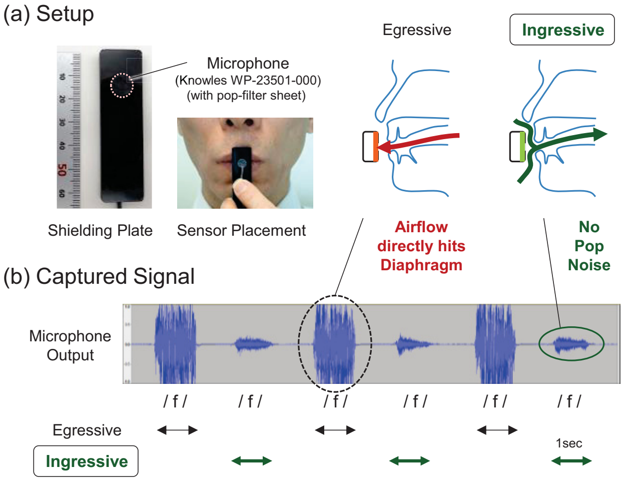


Figure 2: Comparison of pop-noise with egressive and ingressive speech (/f/-sound, microphone distance: 1mm).

In ingressive speech, the microphone can be placed very close to the front of the narrowly opened mouth.

Reducing Pop-Noise

If we want to capture a soft voice efficiently, it is ideal to place the microphone very close to the narrowly opened mouth, because a narrow mouth can concentrate sound energy in a small spot, and sound pressure level attenuates with the square of the distance from the sound source (if a human mouth is regarded as the point source). However, the captured voice may be heavily distorted because expiratory air directly hits the diaphragm of the microphone, especially when uttering the /f/ or /p/ sound (called "pop-noise"). Therefore, the microphone should usually be placed a reasonable distance from the mouth, (or apart from the main air stream); this may cause deterioration of the S/N ratio and increase sound leakage[6].

Figure 2 shows a comparison of captured signals between egressive and ingressive speech of the /f/ sound with a narrowly opened mouth. The shielding plate (18mm width, 63mm height) is placed in front of the mouth with a 1mm gap from the tip of the lips, and a noise canceling microphone (Knowles WP-23501-000) with a pop filter sheet is mounted at the plate with a primary sound port facing the mouth. While repeatedly inhaling and exhaling for 1 second with 1 second intervals for equalizing the amount and speed of airflow for both directions, a captured signal is recorded. This figure indicates that there are severe pop-noises generated in exhaling conditions (equivalent to normal speech and whispering), whereas almost no pop-noises are observed under inhaling conditions (equivalent to ingressive speech) when a similar amount of airflow occurs. Therefore, SilentVoice can put a microphone very close to the front of a narrowly-opened mouth without suffering

[6] Meshed or sponge-like filter materials are usually used as a "pop-guard", however, they are needed to mechanically isolate filter units from the microphone capsule for obtaining good vibration insulation, meaning it is not very effective for small devices such as wearables.

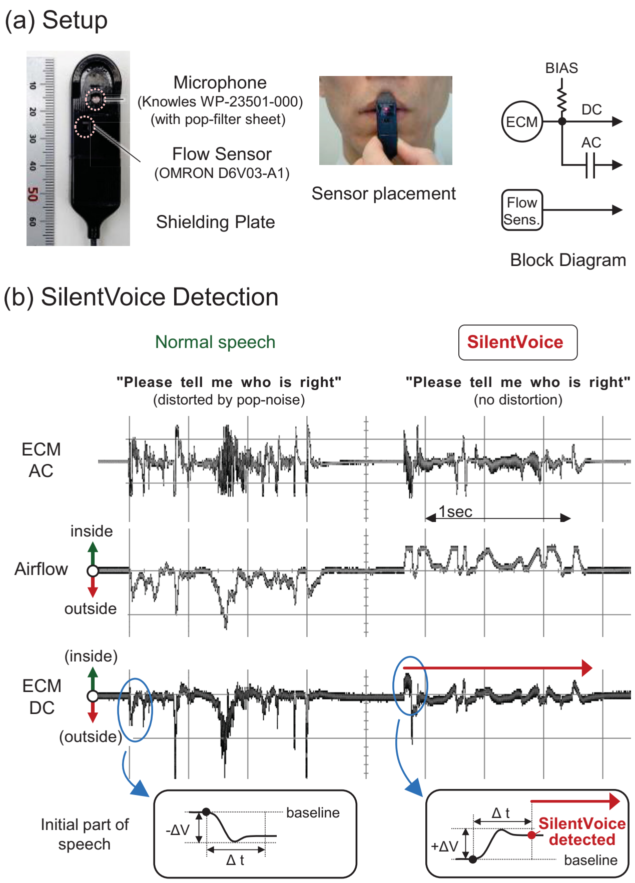


Figure 3: Separation of normal speech and SilentVoice.

SilentVoice can be easily separated from normal speech by measuring airflow direction.

from pop-noise; and can capture a soft voice sound with a good S/N ratio (e.g. if a microphone is placed 2mm from the tip of the lip, about 20dB of signal gain can be achieved compared with mouth-corner placement (about 20mm distance from the tip of the lip)).

Separation of Voice Commands from Normal Conversation

One common problem with the NUI (natural user interface) is how to separate valid commands from our daily actions (sometimes called the "Midas Touch"). Usually, manual switching (e.g. pressing a "talk" button) or uttering activation words (e.g. "Hey Cortana", "Hey Siri", "OK Google", "Alexa", ...) are used for separation in speech input interfaces. SilentVoice's "ingressive speech" rarely occurs in our normal utterances, so that commands can easily be separated from normal conversation by simply measuring airflow directions with flow sensor (e.g. OMRON D6F-V03A1, Figure 3(b, middle)) or pressure sensor, however, such sensors require additional space and cost. On the other hand, typical electret-condenser type microphones (ECM) has sensitivity of low-frequency pressure change [15]. Figure 3(b, bottom) indicates that DC-coupled ECM output reflects airflow, but there is a short time-constant due to ECM's "capacitor"-like structure. However, normal speech and SilentVoice are rarely mixed in single words or sentences. Therefore SilentVoice can be detected based on initial airflow direction of each words or sentences with two threshold values (ΔV and Δt), the detected state is held during consecutive speech is observed.

		Detected	
total:338 pairs		SilentVoice	Normal
Actual	SilentVoice	336	2
	Normal	6	332

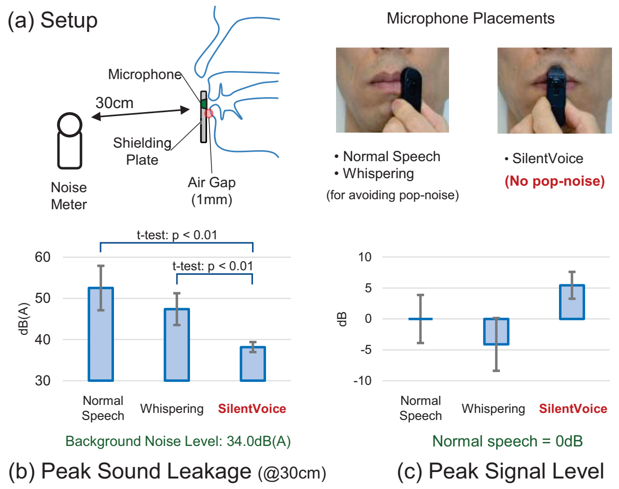
Table 1: Confusion matrix of SilentVoice detection. ΔV:40mV, Δt:5ms. Accuracy = 98.8%

Detection accuracy is evaluated with three adult male subjects (authors not included). Subjects hold a SilentVoice microphone (Figure 3(a)) and repeatedly say "Please tell me who is right." with both normal speech and SilentVoice at least 100 times with 1 second intervals. If the initial part of each sentence has a raised DC level with a 40mV(ΔV) and 5ms(Δt) threshold, it is labeled as SilentVoice. Table 1 shows the confusion matrix for a total of 338 sentence pairs, detection accuracy is 98.8%. In practical conditions, some misdetections may occur when harsh or deep breathing are performed via the mouth, as well as habitual oral-breathing users. However, a speech recognizer can easily ignore these signals because there is no linguistic information contained.

Evaluation of Sound Leakage and Signal Level

Outside sound leakage and the captured signal level of SilentVoice are evaluated along with normal speech and whispering. The shielding plate with microphone (Figure 3(a)) is placed at the corner of the mouth (in both normal speech and whispering conditions to avoid pop-noise), and in front of the mouth (in SilentVoice conditions, no pop-noise) with a 1mm gap from the tip of the lip (microphone diaphragm is placed 2mm from the tip of the lip). Preliminary experiments indicate that the position of the shielding plate does not have any significant effects on the level of sound leakage in both normal speech and whispering conditions. When uttering "Please tell me who is right." in the smallest possible voice, the leaked sound level is measured by a noise meter (UNI-T UT351) from a distance of 30cm with dB(A) and 125msec attack time, with the voice signal simultaneously captured by the noise canceling microphone (Knowles WP-23501-000) embedded in the shielding plate. Seven adult subjects (4 males and 3 females, authors not included) say the sentence 6 times for each condition, and the average peak levels of the leaked sound and voice signal during utterance of each single sentence are calculated. No artificial audio feedback is used such as headphones while speaking. The background noise level of the room is 34.0dB(A).

As shown in Figure 4(b), the averages of peak sound leakage of soft normal speech, soft whispering, and SilentVoice are 52.5dB(A), 47.4dB(A), and 38.2dB(A), respectively. This indicates that the peak leaking sound level of SilentVoice is lower than soft whispering, and below a typical noise level of quiet library conditions (40dB(A)). Here are possible reasons. Basically, it is



(b) Peak Sound Leakage (@30cm)

(c) Peak Signal Level

Figure 4: Comparison of (b) peak sound leakage, and (c) peak captured signal level while uttering "Please tell me who is right." (error bars: standard deviation of the mean).

SilentVoice can achieve low sound leakage (< 40dB(A)) with larger signal level.

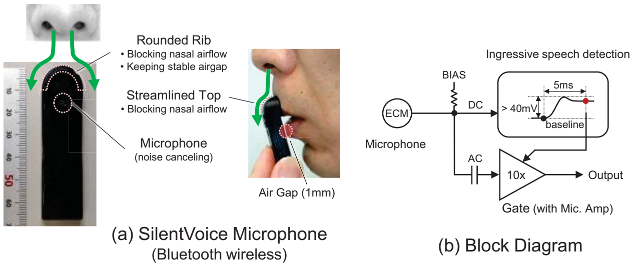
effective in squeezing airflow for decreasing sound leakage. As for normal speech, some "minimum" amount of airflow is needed for vibrating vocal cords. On the other hand, there is no minimum amount of airflow for generating turbulence, and a narrower gap can generate larger turbulence with little airflow; however, we hardly squeeze the gap of our vocal cords while whispering. SilentVoice can provide a narrow and stable air gap by attaching a shielding object.

For further leaked sound reduction, it is effective in providing real-time feedback of captured sound to the user via headphones for maintaining small-but-enough voice volume especially in very quiet (and also very noisy) conditions. Moreover, leaked noise of SilentVoice primarily occurs by "clicking" sounds generated when the tongue or lips are separated (e.g. /t/, /p/). Under very quiet conditions, some surrounding people may be able to hear those sounds, but it is hard to restore the content of the utterance.

Figure 4(c) shows the average peak signal level captured by the microphone. This figure indicates that the signal of SilentVoice can be captured in larger volume than soft normal speech and soft whispering conditions regardless of its small sound leakage because of ultra-closely placed microphones. Therefore, SilentVoice can achieve both low sound leakage and a high S/N ratio.

Microphone Unit

Figure 5 shows a pendant-shaped SilentVoice microphone with a block diagram. The structure is basically similar as a conventional Bluetooth microphone, just adding SilentVoice detection circuit with a 10x microphone amplifier. When SilentVoice is detected, the amplified voice signal is transmitted. An omni-directional type microphone can also be used, but use of a noise canceling



(a) SilentVoice Microphone (Bluetooth wireless)

(b) Block Diagram

Figure 5: (a) SilentVoice microphone, (b) block diagram.

Rounded rib achieves proper air-gap as well as blocking nasal airflow.

(bi-directional) type microphone is strongly recommended for eliminating surrounding noise.

It is effective to keep a narrow air gap between shielding plate and mouth for generating large and substantial turbulence with a small amount of inspiratory air. A large shielding plate can squeeze the air channel and block sound leaks while simultaneously deteriorating portability. The shown "pendant style" prototype is an example design for achieving both good airflow performance and wearability (18mm width, 65mm height). The rounded rib structure of the top position is designed for keeping a narrow gap, microphone position, and clearance from the lower lip during utterance. A stream-lined top shape can also eliminate snort airflow noises.

RECOGNITION OF SILENTVOICE

This section describes how SilentVoice can be recognized with an ordinary voice recognizer by just re-adapting the acoustic model because it retains many sound features of our normal voice.

At first, the recognition rate of SilentVoice is evaluated in combination with a limited dictionary, because one of the initial target applications of SilentVoice is "command-and-control" style voice interaction systems. We select 85 sentences (called a "dataset" hereafter) consisting of typical smartphone operations (e.g. "Check for new mail.", "When next train departs?", "Take a picture.", "Turn on Airplane mode."), and some parameter word sets (e.g. "one, two, ...", "Monday, Tuesday, ...", "January, February, ..."). The dataset is spoken by five adult subjects (3 males and 2 females, authors not included) with SilentVoice, and captured by a SilentVoice sensor unit (Figure 3(a)). The average length of a single dataset (consisting of 85 sentences) is 3.2min.

The Custom Acoustic Model of the Microsoft Bing Speech API [25] [7] is used for evaluation. Three datasets are separately collected for one subject. Two datasets are used for adaptation of a standard acoustic model (The Microsoft Search and Dictation AM), one dataset is used for evaluation of WER (Word Error Rate: (insertions +

[7] The internal structures of recognizer and adaptation methods are not disclosed. A free subscription model is used for some parts of the evaluation.

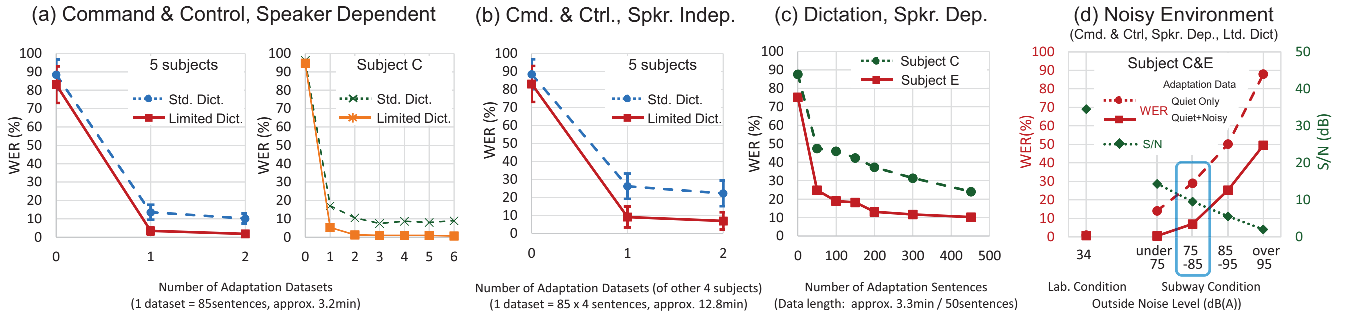


Figure 6: Word Error Rate of SilentVoice (error bars: standard deviation of the mean)

In command & control applications, 1.8% of WER is achieved in speaker-dependent and limited dictionaly conditions after adaptation by 2 sets of 85 command sentences (total adaptation data length: approx. 6.4min).

deletions + substitutions) / total number of words)). A limited dictionary is also created based on a standard language model (The Microsoft Search and Dictation LM) with text data of 85 sentences.

Figure 6(a, left) shows the evaluation results in a speaker dependent condition. Before adaptation, the average WER of 5 subjects is 83.0% (with limited dictionary). It improves to 3.4% and 1.8% after adaptation by one and two datasets, respectively. Therefore, SilentVoice can be used as a "command-and-control" style voice application by collecting at least one dataset of the target user's voice. Figure 6(a, right) shows the long-term adaptation results of subject C with four additional collected datasets. It indicates adaptation by two datasets is enough to settle WER.

However, it is still troublesome to collect each user's voice data before use for practical application. Figure 6(b) shows evaluation results corresponding to speaker independent conditions. An acoustic model is trained by four subjects' combined datasets (the average length of a single combined dataset is 12.8min), and WER is calculated by left one subject's dataset. The process is repeated 5 times by changing subjects for testing, and the average is calculated (leave-one-subject-out). This figure indicates that 7.0% of WER can be achieved after adaptation by two combined datasets without collecting personal voice data before use.

A dictation style application is the final target of SilentVoice. Figure 6(c) shows preliminary evaluation results for unlimited dictionary and speaker dependent conditions. SilentVoice data of TIMIT-SX&SA [11] (452 sentences, average data length is 29.8min), are collected for two adult male subjects (authors not included) for adaptation of a standard acoustic model (The Microsoft Search and Dictation AM); "CID Everyday Sentence List" [7] (List A&B, 20 sentences, average data length is 1.0min) are also collected for WER evaluation with a standard language model (The Microsoft Search and Dictation LM). Before adaptation, the WER for each user is 87.6% (subject C) and 75.2% (subject E), respectively. After adaptation by 452 sentences, WER improves to 24.1% and 10.2%, respectively.

APPLICATIONS

Figure 7 shows some variations of the SilentVoice-embedded devices. We think the primary application is wearable voice-based interaction systems. Tiny devices (e.g. ring and earphone) are not easily noticed by surroundings, and the operation posture of a ring-shaped device (covering the mouth with a handgrip) is very natural. Therefore, the user can confirm schedules or e-mails without annoying other people. SilentVoice can easily be combined with other devices such as SmartWatches, SmartPhones, and TV remotes; a hands-free operation is realized by using an overhead-style headset. Real-time voice communication is also possible by directly transferring whisper-like sound signals such as SilentPhone.

DISCUSSIONS

This section discusses current limitations and possible solutions for SilentVoice.

Outside noise reduction

SilentVoice's ultra-close microphone placement (e.g. 2mm) is effective in reducing outside noise with a noise-canceling type differential microphone. We preliminarily test the S/N ratio and WER of SilentVoice in a subway car; background noise levels are distributed from 61 to 102dB(A), and the average is 81.0dB(A). A single dataset of 85 command sentences (same as previous section, average data length is 2.9min) is recorded by using the SilentVoice microphone with two adult male subjects (Subject C&E, authors not included). Recorded sentences are divided into four groups based on outside noise level. Figure 6(d, green dashed-line,

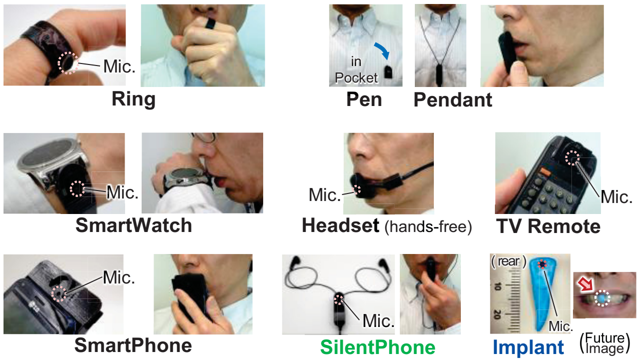


Figure 7: Variations of SilentVoice devices

right scale) shows the S/N ratio in each condition; the leftmost group (34dB(A)) is the quiet condition for reference. It indicates that SilentVoice can be captured with 9.6dB of S/N ratio in about an 80dB(A) noisy environment. WER is also evaluated in a speaker dependent, limited dictionary condition. Figure 6(d, red dashed-line, left scale) shows WER with an acoustic model adapted by two SilentVoice datasets of the subject recorded in quiet conditions (average data length is 5.5min). After additional adaptation by the same subject's other SilentVoice dataset (Subject C: 85 command sentences, total data length is 2.9min; Subject E: "CID Everyday Sentence List (A&B)", 20 sentences, total data length is 0.8min) recorded in the same subway conditions, WER is improved (shown in Figure 6(d, red straight-line, left scale)). It indicates that SilentVoice can be recognized in about an 80dB(A) noisy environment with a WER of 6.9% by re-adaptation with short SilentVoice data under noisy conditions. However, recognition in an over 95dB(A) highly noisy environment is still difficult because the S/N ratio of the captured signal is decreased to almost 0dB.

Recognition of Speaker Independent Dictation

We preliminarily evaluate the WER of SilentVoice under unlimited dictionary and speaker independent conditions. A standard acoustic model (The Microsoft Search and Dictation AM) is trained by two adult male subjects' SilentVoice data (subject C&E, TIMIT-SA&SX, total 908 sentences, 59.6 min, authors not included). Three additional adult female subjects' SilentVoice data ("CID Everyday Sentence List (A&B)", 20 sentences each, average data length is 1.0min, authors not included) are also collected and tested with a standard language model (The Microsoft Search and Dictation LM). Before adaptation, the average WER is 86.4%. After adaptation by the 908 sentences, the average WER remains 73.2%. The result indicates that adaptation with just two users' combined hour-long data is not enough for recognizing dictation type speech in speaker independent conditions (e.g. Several thousand hours of adaptation data is needed for achieving good accuracy for such application [12]).

Phonemes Can (and Cannot) be Input with SilentVoice

SilentVoice uses the same movements of articulation mechanisms (tongue, jaw, lips) as normal speech or whispering, however ON/OFF timing of the sound source (turbulence by airflow) cannot be precisely controlled compared with vocal cords. In whispering and SilentVoice, the sound source is continuously activated while uttering whole sentences. Therefore, the difference between unvoiced consonants and voiced consonants become unclear (e.g. /k/&/g/, /s/&/z/, /p/&/b/) [23]. Nasal sounds (e.g. /m/, /n/) can be generated by ingressive air but are scattered only from the nasal cavity which is far from the microphone, so that the captured signal level may significantly decrease. In addition, the /h/ sound is often missing; the possible reason is the diffused sound energy by stretched lip. Therefore, the corresponding unvoiced and voiced consonants is not clearly separated; the nasal and /h/ sounds are also observed as a <*sil*> (silent) element. Here are some error examples picked up from transcription log (correct → error); (themselves → the self), (the room → through), (here we go → your eagle). For further improvement of recognition accuracy, a SilentVoice-specific pronunciation dictionary is needed; for example ({observed: candidates}), {big: big / pig / pick} (for mixing voiced & unvoiced consonants), {a<*sil*>ber: amber / umber / number / humber} (for missing nasal and /h/ sounds). These phoneme substitutions also occur in real-time voice communication, but our brain can compensate many changes based on context such as talking in whispers.

"Unnoticeability" from Surroundings

There are two kinds of clues that surrounding people can notice SilentVoice operations; leaked voices and operating postures. As for leaked voices, surrounding people may NOT be able to hear leaked voices in many real situations because SilentVoice's leaked voice volume (<40dB(A)) is less than the background noise level. As for operating postures, we think it can be "camouflaged" by commonly-used daily actions such as covering the mouth by a handgrip with a finger-ring shaped microphone.

Learning SilentVoice

SilentVoice's ingressive speech is different from our normal speech manner and requires somewhat unfamiliar body movements for users. Many users can master SilentVoice utterances in less than 15 minutes. However, there are some obstacles; (1) how to shut the nasal cavity when inhaling, (2) how to make a clear SilentVoice sound with a small air flow, and (3) how to utter long sentences.

(1) Inhaling only from the mouth: We usually inhale air from the nose, therefore the nasal cavity is not completely shut while inhaling from the mouth. Using a "straw" is effective for learning how to inhale only from the mouth.

(2) Making a clear SilentVoice: Basically, squeezing the air channel is effective for generating large turbulence and making a clear SilentVoice sound with a small airflow. Narrowly opening the mouth like in "ventriloquism" is a practical method. Closely placing the microphone is necessary not only for making a narrow air channel but also for maximizing S/N.

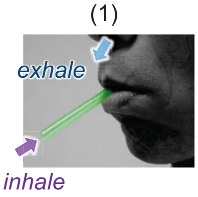
(3) Uttering long sentences: While there are no problems for short word inputs with SilentVoice, users often encounter "cannot inhale anymore" conditions when uttering longer sentences. In normal speech, we usually start an utterance after inhaling. In the case of SilentVoice, we should "exhale" before an utterance, especially for speaking long sentences. However, we (unconsciously) inhale before utterance, and there is no capacity left in our lungs for SilentVoice. It is effective to consciously exhale before speaking long sentences to prevent "full lung" conditions[8].

Detailed instruction is shown in the Appendix.

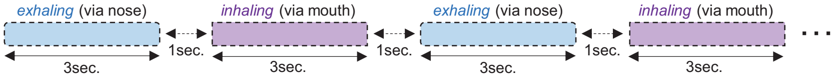
[8] Well-trained users can utter speech with SilentVoice for more than 10 seconds with one breath (e.g. Subjects C&E).

[How to Learn **SilentVoice**]

[Step 1] Mastering the breathing method (how to inhale only from the mouth, not from the nose)



(1) Lightly hold a straw with the lips.
Slowly *inhale (breathing-**in**)* **as if sucking through a straw**, *exhale (breathing-**out**)* through the nose.
Do not *inhale* from the nose. Try to take a long breath by using slow **abdominal respiration**.



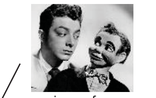
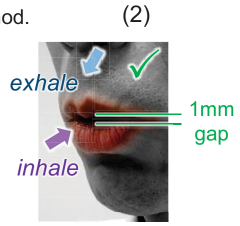


image from
ventriloquistcentral.com

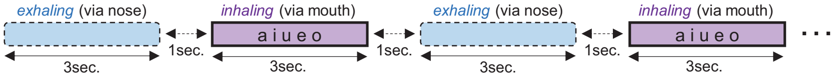
(2) Remove the straw, relax the jaw, gently open the lips with a small gap (e.g. 1mm) as performing "ventriloquism".
Repeat (1) as if **sucking through a (virtual) straw**. **Do not *inhale* from the nose**.
You may hear turbulence sound while *inhaling*, find the proper lip shape that makes a large sound with a small breath.
Place the finger in front of the nasal hole.
If you don't feel any airflow at the nasal hole while *inhaling*, congratulations! You have mastered the breathing method.

[Step 2] Mastering SilentVoice articulation (how to move the mouth while inhaling)

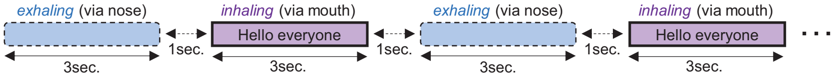


relax jaw
slightly open the lips

(3) Repeat (2), making same lip, jaw, and tongue movements as if whispering " **a i u e o** " while ***inhaling*** .
You may hear whisper-like " **a i u e o** " sound while *inhaling*.
Try to keep lip shape restricted to a 1mm gap while uttering.
Do not vibrate your vocal cords, do not make whisper sounds from deep in the throat.
Find the best lip, jaw, and tongue movements that make large and clear sounds with small breaths.



(4) Repeat (3), change the lip, jaw, and tongue movements as if whispering "**Hello everyone**".
You can close the lips when uttering /v/ (and also /p/, /b/, /w/, /m/, and /n/) for improving sound quality.
Find the best lip, jaw, and tongue movements that make large and clear sounds with small breaths.



(5) If you can hear a whisper-like "Hello everyone" while *inhaling*, congratulations! You have mastered SilentVoice.

Appendix: SilentVoice instruction sheet for subjects.
All subjects are trained before testing. The typical training time is 15 minutes.

Sampling Rate

SilentVoice has much richer high-frequency components compared with normal speech and whispering (Figure 1). Therefore, use of a higher sampling rate (at least 16kHz) for voice recognition is recommended. For the same reason, VoLTE networks (50-7kHz bandwidth) are recommended for real-time voice communication with cellular phones.

Speaking aid application for laryngectomees

SilentVoice was originally developed for normal people's daily use. There exists the possibility of using it as a speaking aid system for (total) laryngectomy patients. An issue that arises when using SilentVoice for laryngectomees is the inability to breathe via mouth due to a separated trachea. In esophageal speech (that uses the esophagus as proxies for vocal cords and lungs), air is injected into the esophagus by abdominal respiration; this method may be used with SilentVoice. It still requires some training, however, as SilentVoice can generate source sounds much easier than esophageal oscillation, meaning it can be used as an alternative speaking aid.

CONCLUSION

This paper proposes SilentVoice, which enables voice input with very low sound leakage (less than 39dB(A)) by using the proposed "ingressive speech" method. Since pop-noise does not occur, the microphone can be placed very close to the front of the mouth (less than 2mm), and can capture an ultra-small speech sound with a good S/N ratio. Normal speech and SilentVoice can easily be separated with 98.8% accuracy by measuring initial airflow direction when an utterance occurs. Recognition of SilentVoice is also possible with a specially trained acoustic model. The current WER is about 1.8% (in speaker dependent conditions), and 7.0% (in speaker independent conditions) with a limited dictionary of 85 command sentences. S/N ratio and WER in noisy condition are also confirmed. It indicates that SilentVoice can be used for command-and-control style voice interaction systems in not only quiet spaces but also noisy environments. As for dictation style applications, it still requires each user's half-hour SilentVoice data before use. For realizing practical speaker-independent system, massive scale data collection is needed along with creating SilentVoice-specific pronunciation dictionary. An effective training method should be established for spreading new speech "skill" to the public. Field tests including social acceptability are also needed for practical applications. In the future, it is possible that we (i.e. humans) have the option of augmentation by implanting tooth-shaped SilentVoice devices (Figure 7(bottom-left): mockup image[9]), so that all conversations (including face-to-face communication) will be performed via SilentVoice with a secure network. Our world will become much quieter.

[9] Microphone port is placed at rear part of the teeth for avoiding covered by the lip.

REFERENCES

1. Amazon Alexa. Retrieved July 30th, 2018 from https://developer.amazon.com/alexa
2. ambient::technology. Retrieved July 30th, 2018 from http://wayback.archive.org/web/20150330030801/ http://www.theaudeo.com/?action=technology (web archive)
3. Apple Siri. Retrieved July 30th, 2018 from http://www.apple.com/iphone/features/siri.html
4. A. Bedri, H. Sahni H, P. Thukral, T. Starner, D. Byrd, P. Presti, and Z. Guo. 2015. Toward Silent-Speech Control of Consumer Wearables. IEEE Computer, 48(10), 54-62. https://doi.org/10.1109/MC.2015.310
5. W. Chan, N. Jaitly, Q. Le, and O. Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Proc. of ICASSP, 4960-4964. https://doi.org/10.1109/ICASSP.2016.7472621
6. M. H. Cohen, J. P. Giangola, and J. Balogh. 2004. Voice user interface design. Addison-Wesley Professional, 2004.
7. H. Davis, and S. Silverman. 1970. Hearing and Deafness. Holt, Rinehart, and Winston (New York).
8. B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg. 2010. Silent speech interfaces. Speech Communication, 52(4), 270-287. https://doi.org/10.1016/j.specom.2009.08.002
9. R. Eklund. 2008. Pulmonic ingressive phonation: Diachronic and synchronic characteristics, distribution and function in animal and human sound production and in human speech. Journal of the International Phonetic Association, 38(03), 235-324. https://doi.org/10.1017/S0025100308003563
10. M. L. Fouquet, A. J. Gonçalves, and M. Behlau. 2009. Relation between videofluoroscopy of the esophagus and the quality of esophageal speech. Folia Phoniatrica et Logopaedica, 61(1), 29-36. https://doi.org/10.1159/000191471
11. J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. 1986. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NIST Interagency/Internal Report (NISTIR), 4930.
12. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine, 29(6), 82-97. https://doi.org/10.1109/MSP.2012.2205597
13. T. Hueber, E. L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone. 2010. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. Speech Communication, 52(4), 288-300. https://doi.org/10.1016/j.specom.2009.11.004
14. Hushme. Retrieved July 30th, 2018 from http://gethushme.com/
15. T. Ishigaki. 2005. Development of Two Measurement Systems of Abnormal Condition by Using Condenser Microphone. Master Thesis, Hosei University (in Japanese).
16. T. Itoh, K. Takeda, and F. Itakura. 2001. Acoustic analysis and recognition of whispered speech. In Proc. of ASRU'01, 429-432. https://doi.org/10.1109/ASRU.2001.1034676
17. S. C. S. Jou, T. Schultz, and A. Waibel. 2004. Adaptation for soft whisper recognition using a throat microphone. In Proc. of ICSLP2004, 527-530.
18. A. Kapur, S. Kapur, and P. Maes. 2018. AlterEgo: A Personalized Wearable Silent Speech Interface. In Proc. of IUI 2018, 43-53. https://doi.org/10.1145/3172944.3172977
19. J. J. Li, I. V. McLoughlin, L. R. Dai, and Z. H. Ling. 2014. Whisper-to-speech conversion using restricted Boltzmann machine arrays. Electronics Letters, 50(24), 1781-1782. https://doi.org/10.1049/el.2014.1645
20. H. Manabe, A. Hiraiwa, and T. Sugimura. 2003. Unvoiced speech recognition using EMG-mime speech recognition. In Ext. Abst of CHI'03, 794-795. https://doi.org/10.1145/765891.765996
21. K. Mase, and A. Pentland. 1991. Automatic lipreading by optical - flow analysis. Systems and Computers in Japan, 22(6), 67-76. https://doi.org/10.1002/scj.4690220607
22. M. Matsumoto, and J. Hori. 2014. Classification of silent speech using support vector machine and relevance vector machine. Applied Soft Computing, 20, 95-102. https://doi.org/10.1016/j.asoc.2013.10.023
23. G. N. Meenakshi, and G. P. Kumar. 2015, A discriminative analysis within and across voiced and unvoiced consonants in neutral and whispered speech in multiple indian languages. In Proc. of INTERSPEECH2015, 781-785.
24. Microsoft Cortana. Retrieved July 30th, 2018 from https://www.microsoft.com/en-us/windows/cortana
25. Microsoft Custom Speech Service. Retrieved July 30th, 2018 from https://azure.microsoft.com/en-us/services/cognitive-services/custom-speech-service/
26. Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In Proc. of ICASSP'03, 5, V-708. https://doi.org/10.1109/ICASSP.2003.1200069

27. K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano. 2006. Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech. In Proc. of INTERSPEECH2006. http://hdl.handle.net/10061/8140
28. (for example) Nuance. Retrieved July 30th, 2018 from https://www.nuance.com/
29. J. Reighard, H. S. Jennings. 1951. Anatomy of the cat. Henry Holt and Company.
30. A. D. Rubin, V. Praneetvatakul, S. Gherson, C. A. Moyer, and R. T. Sataloff. 2006. Laryngeal hyperfunction during whispering: reality or myth? Journal of Voice, 20(1), 121-127. https://doi.org/10.1016/j.jvoice.2004.10.007
31. G. Saon, T. Sercu, S. J. Rennie, and H-K. J. Kuo. 2016. The IBM 2016 English Conversational Telephone Speech Recognition System. In Proc. of INTERSPEECH2016, 1460. https://doi.org/10.21437/Interspeech.2016-1460
32. V. A. Tran, G. Bailly, H. Lœvenbruck, and T. Toda. 2010. Improvement to a NAM-captured whisper-to-speech system. Speech communication, 52(4), 314-326. https://doi.org/10.1016/j.specom.2009.11.005
33. D. Wigdor, and D. Wixon. 2011. Brave NUI World: Designing Natural User Interfaces for Touch and Gesture (1st ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
34. W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. 2017. The Microsoft 2016 conversational speech recognition system. In Proc. of ICASSP, 5255-5259. https://doi.org/10.1109/ICASSP.2017.7953159