Silent-speech enhancement using body-conducted vocal-tract resonance signals

Tatsuya Hirahara [1], Makoto Otani [1], Shota Shimizu [1],
Tomoki Toda[2], Keigo Nakamura[2], Yoshitaka Nakajima[2], Kiyohiro Shikano[2]

1: Toyama Prefectural University,
Department of Intelligent Systems Design Engineering.
5180 Kurokawa, Imizu, Toyama 939-0398, Japan

2: Nara Institute of Science and Technology,
Graduate School of Information Sciences.
8916-5 Takayama, Ikoma, Nara 630-0192, Japan

Corresponding author: Tatsuya Hirahara
Toyama Prefectural University
Department of Intelligent Systems Design Engineering
5180 Kurokawa, Imizu, Toyama 939-0398, Japan
Tel: +81 766-56-7500 ext 459,
Fax: +81 766-56-8030
E-mail: hirahara@pu-toyama.ac.jp

Abstract

The physical characteristics of weak body-conducted vocal-tract resonance signals called non-audible murmur (NAM) and the acoustic characteristics of three sensors developed for detecting these signals have been investigated. NAM signals attenuate 50 dB at 1 kHz; this attenuation consists of 30-dB full-range attenuation due to air-to-body transmission loss and −10 dB/octave spectral decay due to a sound propagation loss within the body. These characteristics agree with the spectral characteristics of measured NAM signals. The sensors have a sensitivity of between −41 and −58 dB [V/Pa] at 1 kHz, and the mean signal-to-noise ratio of the detected signals was 15 dB. On the basis of these investigations, three types of silent-speech enhancement systems were developed: (1) simple, direct amplification of weak vocal-tract resonance signals using a wired urethane-elastomer NAM microphone, (2) simple, direct amplification using a wireless urethane-elastomer-duplex NAM microphone, and (3) transformation of the weak vocal-tract resonance signals sensed by a soft-silicone NAM microphone into whispered speech using statistical conversion. Field testing of the systems showed that they enable voice impaired people to communicate verbally using body-conducted vocal-tract resonance signals. Listening tests demonstrated that weak body-conducted vocal-tract resonance sounds can be transformed into intelligible whispered speech sounds. Using these systems, people with voice impairments can re-acquire speech communication with less effort.

Keywords: non-audible murmur, body-conducted sound, voice conversion, talking aids,

1. Introduction

While microphones have been used in many scientific fields to sense speech sounds, a recently developed device called a "non-audible murmur (NAM) microphone" is receiving increasing attention as a new means for picking up body-conducted speech (Heracleous *et al.*, 2003, Nakajima *et al.*, 2006, Toda *et al.*, 2005b, Nakamura *et al.*, 2006). Typically, the speech sound used is air-borne sound, which is small, fast vibration of the air. Vibration of the air column in the vocal tract vibrates the tract wall, and some of the sound energy generated passes through the tissues of the neck and chest. The body-conducted sound that travels through the neck tissue can be sensed using a sensor modified from a microphone. Actually, speech sounds propagate not only through the air and the bone but also through the body tissue, including the muscles. Nakajima, a pioneer in NAM development, found that murmured speech, which is usually unheard by people nearby, can be detected by using a sensor attached to the neck behind the ear (Nakajima *et al.*, 2003a). This body-conducted weak murmur is called "non-audible murmur (NAM)", and the sensor is known as a NAM microphone.

Nakajima originally detected NAM using a stethoscopic NAM microphone (Nakajima *et al.*, 2003a, 2003b), which is an electret condenser microphone (ECM) implanted into a standard medical-use stethoscope with the tubes removed. When this device is placed on the skin under the mandible, weak whispers inaudible to nearby listeners and even the speaker can be sensed. The optimal position for sensing the NAM signal is the lower part of the mastoid process, i.e. high on the neck behind the ear.

The bandwidth of the NAM signals sensed by the stethoscopic NAM microphone is limited to about 2 kHz because the air between the vibration plate of the stethoscope and the diaphragm of the ECM affects the sensitivity of skin vibration capture. To improve the frequency response, a soft-silicone (SS) NAM microphone was developed that does not have this air cavity (Nakajima, 2005, Nakajima *et al.*, 2005, 2006). This NAM microphone has a simple structure; the exposed ECM diaphragm is covered with soft silicone to form a rigid cylindrical case. Since the acoustic impedance of soft silicone closely matches that of the skin, the soft-tissue vibration directly vibrates the diaphragm with minimal loss. As a result, the SS NAM microphone has a broader frequency bandwidth than the stethoscopic NAM microphone.

The SS NAM microphone, however, is large and heavy, so a neckband is needed to hold it in place. It generates rubbing noise as the user moves; and the bandwidth is still narrower than that of ordinary microphones. Moreover, there are unanswered questions about the propagation characteristics of the vocal-tract resonance sound through the body, the physical characteristics of NAM microphones, and the acoustic characteristics of the NAM signal itself. These drawbacks and unanswered questions have limited the application of NAM microphones.

Nevertheless, the use of body-conducted speech, by means of a NAM microphone in particular, would expand the range of speech technologies. As NAM sounds are imperceptible, NAM microphones can be applied to automatic speech recognition (ASR) systems and telephone systems that require privacy in open situations. As NAM microphones are more sensitive to body-conducted sounds and less sensitive to air-borne sounds, the use of such microphones can improve the noise tolerance of ASR systems (Heracleous *et al.*, 2003, Nakajima *et al.*, 2006). In addition, as NAM microphones are capable of detecting very weak body-conducted speech signals, a weak, virtually imperceptible sound source is enough to activate the microphone without whispering aloud. This sensitivity suggests that externally induced weak vibration of the neck skin can be used to produce laryngeal speech with an electrolarynx (EL) with minimal stray noise and low-energy consumption by the vibrator. Thus, the silent speech technology using NAM microphone is well suited for use in speaking aids for voice impaired people and in communication aids for laryngectomized patients as an alternative to the conventional EL or esophageal speech.

Voice impairments can prevent people from performing daily activities that require verbal communication. When vocal function is severely degraded due to organic or neuromotor disease or even lost as a result of a surgical procedure such as laryngectomy, the vocal source for speech communication is lost. This causes a breakdown in speech communication, the major medium of communication among people. Losing the ability to vibrate one's vocal folds makes it greatly difficult to have a family discussion, to make phone calls, to work with other people, and to perform

other everyday activities.

Alternative methods for speech production without vocal-fold vibration are the use of an EL or of a pseudo glottis with an esophageal airflow. A big problem with EL speech is intense noise from the vibrator; another is monotone speech due to the difficulty of controlling pitch. To these problems, researchers have worked on removing the intense noise made by EL devices (Espy-Wilson *et al.*, 1998) and on developing pitch-controllable ELs (Uemi *et al.*, 1994, Kikuchi and Kasuya, 2004). The battery dependency and one-hand-occupied operation also restrict the user's social activities. While esophageal speech is not dependent on a device or hand operation, it does take a long time and the help of a speech therapist to become proficient in this method. Although both methods enable voice impaired people to speak, the produced speech sounds are of poor quality in terms of naturalness and intelligibility and lack individual vocal characteristics. Despite these deficits, a number of voice impaired people have been using these methods while waiting for better speaking aids to become available.

We have developed three types of speaking aid systems using three key technologies: high-sensitivity sensing for detecting body-conducted sound, wireless transmission, and voice-quality transformation for speech output. In each one, the sensor is attached to the neck and detects weak vocal-tract resonance sound caused by weak breathing noise or by a small low-amplitude vibrator placed on the neck. In the most promising of the systems, the sound is transformed into intelligible whispered speech by means of a voice-quality conversion algorithm. Using these systems, people with voice impairments can re-acquire speech communication with less effort.

This paper describes the nature of the weak vocal-tract resonance signals in Section 2, the measurement of the acoustic characteristics of three NAM sensors in Section 3, the acoustic characteristics of NAM signals in Section 4, and three systems for enhancing silent speech using weak vocal-tract resonance sounds in Section 5.

2. Nature of weak vocal-tract resonance signals

While development of the NAM microphone has improved the detection of body-conducted speech sounds with a stethoscope, the physical nature of body-conducted speech sounds has been left underexplored. We investigated the process of vocal-tract resonance sounds being conducted through the body and being detected by a NAM microphone placed on the neck.

2.1 Numerical simulation

Sound originating in the vocal tract propagates from the air column inside the tract to the body tissue and can be detected at the body surface. The shortest propagation distance from the vocal tract to the neck surface, where a NAM microphone is commonly placed, is approximately 70 mm. To clarify the NAM transfer characteristics, we numerically investigated sound propagation from the vocal tract to the neck surface using a two-dimensional finite difference time domain (FDTD) method and a head model constructed on the basis of magnetic resonance imaging (MRI) scans.

Three-dimensional geometrical data of a human head during production of the vowel sound /e/ were obtained using phonation-synchronized MRI scans (Nota *et al.*, 2005). An image on the midsagittal plane was extracted to produce a two-dimensional head model. For simplicity, a homogeneous head model was generated, i.e. the head was approximated as being composed of only soft tissue. Finally, the vocal tract was replaced by a simplified model with a rectangular duct with a cross-section 30 mm wide. This approximation of the vocal-tract shape was made because a scanned vocal tract is extremely narrow in the vicinity of the larynx, and it would require a grid geometry that would be too fine to be simulated by the FDTD method. The physical parameters of the soft tissue were set in accordance with a published report (Oestreicher, 1951). Cyber Logic Wave 2000 Pro was used as the FDTD solver. Details of the simulation are described elsewhere (Otani *et al.*, 2009).

2.2 Simulation results

Figure 1(a) shows the geometry of the simulated region including the head, the sound source, and the receivers. The sound source was located in front of the mouth, assuming a reciprocal theorem, as

was a pulse-driven vibrating plate. Receivers were located at the inlet and outlet of the vocal tract and on the neck behind the ear, where a NAM microphone would normally be attached. These receivers are referred to as "*in*", "*out*", and "*nam*", respectively. The sound intensity levels at each receiver I_{in}, I_{out}, and I_{nam} were calculated, and then I_{in}/I_{out} and I_{nam}/I_{out} were calculated as shown in Figure 1(b). The I_{in}/I_{out} corresponds to the vocal-tract transfer functions, showing spectral peaks, i.e. formants, at 0.5, 1.4, 2.1, and 3.0 kHz. The $I_{nam}/I_{out} = (I_{nam}/I_{in}) / (I_{out}/I_{in})$ corresponds to the sound intensity level observed at *nam* assuming that the sound radiated from the mouth has flat frequency characteristics. The transmission loss of sound as it passed from the air in the vocal tract into the soft tissue and the propagation loss through the soft tissue $I_{nam}/I_{in} = (I_{nam}/I_{out}) / (I_{in}/I_{out})$ were both approximately −50 dB at 1 kHz. The loss consists of 30-dB full-range attenuation due to air-to-soft-tissues transmission loss and a −10 dB/octave spectral decay due to propagation loss in the soft tissues. The peaks at 6.3 and 9.7 kHz are attributed to the eigenfrequencies of the head (Fujisaka *et al.*, 2004).

Insert Fig. 1 (a), (b) here

---

3. Measurement of acoustic characteristics of NAM microphones

While NAM microphones have been improved due to the efforts of many researchers, their acoustic characteristics have been left underexplored. We measured the acoustic characteristics of three types of NAM microphones.

3.1 Specifications and structures of microphones

As mentioned, a NAM microphone has an electret condenser microphone with an exposed diaphragm that is covered with a soft polymer material, such as soft silicone, so as to provide better impedance matching with the soft tissue of the neck. To overcome the problems of the SS-type NAM microphone mentioned in the introduction, we built two new types of NAM microphone. They are smaller than the SS-type microphone. Urethane elastomer is used rather than soft silicone because it adheres to the skin. These key design changes result in a light unit that does not need to be held in place with a neckband and minimal generation of rubbing noise, which helps overcome the impediments to practical use of NAM microphones.

The specifications and structures of three types of NAM microphone are shown in Table 1 and Figure 2. The SS-type microphone is manufactured by Mitsumi Electric Co., Ltd; an ECM is covered with soft silicone and placed in a rigid cylindrical case 30 mm in diameter and 20 mm high. The unit weighs 27 grams. Figure 3(a) shows an x-ray computed tomographic image of an SS NAM microphone. The ECM is located at a very shallow position. The distance between the surface and the ECM diaphragm is about 1 mm. The urethane-elastomer (UE)-type NAM microphone we developed has an ECM covered with urethane elastomer and is placed in a rigid cylindrical case 20 mm in diameter and 10 mm high. The unit weighs 5.5 grams. A metal corn reflector is placed below the ECM, and the lower part of the reflector is filled with hard silicone. The urethane elastomer is adhesive itself, which facilitates the attachment of the NAM microphone to the skin. The urethane-elastomer-duplex (UED)-type NAM microphone has an ECM covered with urethane elastomer and is placed in a rigid coaxial cylindrical case 8.5 mm in diameter and 6.5 mm high, which in turn is placed in a rigid coaxial cylindrical case 17 mm in diameter and 8 mm high. The unit weighs 3 grams. The urethane elastomer filling the central trench around the ECM acts as a medium through which sound is carried, and the urethane elastomer filling the outer trench acts as an adhesive, facilitating attachment of the microphone to the skin and serving as an insulator against air-conducted sound. Figure 3(b) shows an X-ray computed tomographic image of a UED NAM microphone. The ECM is again located in a very shallow position. The distance between the surface and the ECM diaphragm is about 2.5 mm.

Insert Table 1, Fig. 2 and Fig.3 (a),(b)here

3.2 Frequency characteristics of NAM microphones

The frequency characteristics of the NAM microphones were measured using an audio analyzer (Pulse 3560C; Brüel & Kjær), an accelerometer (NP-3211; Ono Sokki Co., Ltd.), a bone conduction vibrator (BR-41; Rion Co., Ltd.), and a urethane elastomer cylinder 75 mm in diameter and 50 mm high, which simulates the soft tissues in the neck.

Figure 4 shows a diagram of the system used to calibrate the NAM microphones. The vibrator was placed beneath the cylinder. A NAM microphone and the accelerometer were placed on the top of the cylinder. The accelerometer was covered with a silicone enclosure to align its height to that of the NAM microphone. A 1-kg weight was placed on top of the silicone enclosure to stabilize the conditions of contact between the microphone and cylinder and between the accelerometer and cylinder. The vibrations transmitted through the cylinder were sensed by the microphone and the accelerometer, and these signals were compared analytically.

Insert Fig.4 here

Let the measured frequency response of the accelerometer be A(ω); then A(ω)/ω2 gives the frequency response of the displacement. The accelerometer had a flat acceleration response at frequencies up to 10 kHz. Since NAM microphones are displacement sensors, comparison of the calibrated frequency response of the microphone, D(ω), with A(ω)/ω2 yielded the frequency response of the microphone. The frequency response of the NAM microphone amplifier was cancelled.

Figure 5 shows the mean frequency response for the three types of NAM microphones: three SS ones, three UE ones, and a UED one. The mean frequency response for the SS ones resembles that of the UE ones at frequencies below 2 kHz, except that the sensitivities were different at 1 kHz. The UE response had no peak at approximately 2 kHz whereas the SS and UED responses exhibited a peak. At frequencies less than 0.4 kHz, both the SS and UE microphones had lower sensitivities. The SS frequency response was lower at approximately 8 kHz, whereas the UE response rose gradually by 6 dB/octave above 1.5 kHz. The SS sensitivities at 1 kHz were between −44 and −51 dB [V/Pa]; those of the UEs at 1 kHz were between −44 and −58 dB [V/Pa]. The UED one had the highest sensitivity at frequencies less than 1.2 kHz. The frequency response of the UED one resembles that of the SS ones between 1.4 and 4 kHz and resembles that of the UE ones at frequencies greater than 4 kHz. The sensitivity of the UED microphone at 1 kHz was −41 dB [V/Pa].

Insert Fig.5 here

---

4. Acoustic characteristics of NAM signal

We investigated the acoustic characteristics of the NAM signals recorded with the three types of NAM microphones. We also compared the acoustic characteristics of signals recorded with a UED NAM microphone with those recorded with an SS NAM microphone.

4.1 Recording procedure

The NAM signals were recorded in a soundproof room. Six male and seven female adult participants read 50 ATR phoneme-balanced Japanese sentences (Abe *et al.*, 1990, Sagisaka and Uratani, 1992) in NAM mode of weak whispering. An SS NAM microphone was used. It was attached to the neck behind the ear below the mastoid process and fixed in place with a neckband. The microphone was carefully positioned so as to optimize detection of the NAM signals. The signals were recorded using a solid state recorder (PMD670; Marantz) at a sampling rate of 48 kHz with 16-bit resolution. During the recording, the NAM phonation mode and the signal level were monitored by an audio expert to maintain the quality of the recorded signals. When intense noise was detected, the speaker was instructed to reread the sentence. In total, 80-minute NAM signals for 640 sentences were recorded.

4.2 Waveform and spectrogram of NAM signals

The waveforms and spectrograms of two example recorded signals are shown in Figure 6. A male and female speaker each read a Japanese sentence, “arajuru geN□□□...。□...□□.□□”; its meaning is "all the reality". Despite the low-amplitude nature of the NAM signals, speech segments are identifiable in the waveform. The background noise level, however, was high, and a number of impulsive pop noises overlapped the NAM signals, suggesting that the signal-to-noise ratio (SNR) of the signals was low.

Insert Fig.6 (a), (b) here

The spectrograms also suggest that the SNR was low and that the bandwidth of the signal was narrow. Formants are seen during vowels, liquids (/r/) and palatal approximant (/ju/) parts in the spectrograms. Wide-band impulsive noises, although their bandwidths are limited within a few kHz, at plosives (/tsu/ at 1.28 s in Fig. 6 (a) and at 1.62 s in Fig. 6 (b), /t/ at 2.36 s in Fig. 6 (a) and at 2.98 s in Fig. 6 (b)) and fricatives (/□□/ at 1.11 s in Fig. 6 (a) and at 1.43 s in Fig. 6 (b)) are also prominent in the spectrograms. On the other hand, impulsive noise at voiced plosives (/b/ at 2.15 s in Fig. 6 (a) and at 2.77 s in Fig. 6 (b)) are not seen in the spectrograms. Energy at fricatives (/s/) is not clearly seen at around 1.9 s in Fig. 6(a), but it is seen at around 2.4 s in Fig. 6 (b). The content of the utterance was clearly audible for the male speaker. Although the formant pattern of the female NAM signal was vague, the content of the utterance was nevertheless also audible.

The pop noises were generated either by contact between the microphone and skin or by contact of the articulatory organs when producing plosives and fricatives. Some of this type noise can be reduced by level thresholding as the amplitude of the noise was generally larger than that of the signals. After such pop noises are reduced, the speech segments of NAM signals can be automatically detected using their power and delta-cepstrum coefficients.

4.3 SNR of NAM signals

The SNRs of the NAM signals were significantly lower than those of ordinary speech signals obtained with a regular microphone. The mean SNR was 15 dB for NAM signals recorded at a sampling rate of 48 kHz. Reducing the sampling rate improved the SNR—it reached a maximum of 21 dB at a sampling rate of 2 kHz.

4.4 Long-term average spectra of NAM signals

The long-term average spectra (LTAS) of NAM signals resampled at 16 kHz are shown in Figure 7: the black and gray solid lines indicate the mean LTAS for six male and seven female speakers, respectively, and the dotted line shows the noise floor. The energy concentration was found to be around 500 to 800 Hz for both the male and female speakers. Their spectral roll-off above 800 Hz was at −17 dB/octave and that below 500 Hz was at 10 dB/octave. The LTAS data show that the bandwidth of the signals was 3 kHz. The noise floor level of the recording system and the NAM signal strength limited the bandwidth.

Insert Fig.7 here

The frequency responses of the NAM microphone as well as the NAM microphone amplifier played a role in determining the overall frequency characteristics of the NAM signals. By subtracting those frequency responses from the LTAS shown in Figure 7, we obtained the LTAS of underlying NAM signal, as shown in Figure 8. The LTAS show an energy concentration around 200 to 300 Hz, and the spectrum rolls off at −23 dB/octave above 300 Hz. According to the source-filter theory of speech production, the LTAS of ordinary voice has an approximately −6 dB/octave roll-off. As the radiation characteristic from the lips is +6 dB/octave, the spectral tilt of the source was estimated to be −12 dB/octave for ordinary voice. Thus, the underlying NAM signal involves a −11 dB/octave spectral decay due to voice sound transfer through the body. The characteristics of this transfer are roughly equal to the spectral decay suggested by the numerical simulation described in section 2, i.e. −10 dB/octave.

Insert Fig.8 here

4.5 Comparison of NAM signals recorded with SS and UED microphones

Figures 9(a) and (b) shows the waveforms and spectrograms of NAM signals recorded simultaneously using SS and UED NAM microphones. The sentence spoken was “youkoso tojama keNritsu daigaku e”; the meaning is "Welcome to Toyama Prefectural University”. For this recording, the NAM microphones were attached to both sides of a male speaker's neck. Pop noises overlapped the NAM signal recorded with the SS microphone from 0 to 1 s. These noises were less prominent in the signal recorded with the UED microphone. For the SS-recorded signal, the formant pattern is clear at frequencies below 3 kHz, but it is only barely visible above 3 kHz. The spectrogram of the SS-recorded signal was limited to frequencies less than 3 kHz. In contrast, the formant pattern of the UED-recorded signal is clear up to 6 kHz although it is weak in amplitude. The UED-recorded signal had a wider spectrum than that of the SS-recorded one. The frequency responses of both microphones, as presented in Figure 6, are well reflected in the spectrograms.

Insert Fig.9 (a),(b) here

---

5. Three systems for enhancing silent speech using weak vocal-tract resonance sounds

Subsequent to our investigation of the physical nature of NAM signals and the acoustic characteristics of NAM sensors, we built and tested three types of silent-speech enhancement systems: (1) simple, direct amplification of weak vocal-tract resonance signals using a wired UE NAM microphone, (2) simple, direct amplification using a wireless UED NAM microphone, and (3) transformation of the weak vocal-tract resonance signals sensed by a wired SS NAM microphone into whispered speech using statistical conversion.

5.1 Type 1: simple amplification

The first type simply uses amplification of the weak vocal-tract resonance signals, and the system comprises a small electric vibrator, a UE NAM microphone, an amplifier, and a loudspeaker. The vibrator, either an electrodynamic or piezoelectric one, is attached to the neck and transmits weak vibration to the throat. The vibration is weak enough so that nearby listeners are not disturbed by stray sounds. The small vibration of the tissue created by the vibrator reaches the air column of the vocal tract and drives a weak vocal-tract resonance that changes with the behavior of the articulators. These resonance sounds then propagate through the body and vibrate the skin surface of the neck. The microphone picks up the small vibration at the neck surface, and the picked-up signal is amplified and transduced into sounds by the loudspeaker.

5.2 Type 2: simple amplification with wireless transmission

The second type is the same as the first but uses a wireless UED NAM microphone and a wireless receiver. The microphone is free of the noises caused by a microphone cable contacting the wearer’s clothing. Figure 10(a) shows a UED NAM microphone embedded in a Bluetooth transmitter which uses an advanced audio profile. The Bluetooth transmitter is attached to the neck behind the speaker’s ear, and the listener receives the NAM signal with a Bluetooth headset designed for cell phones. Various Bluetooth audio adaptors and audio receivers available in the market can be used. The transmission range is up to 9 m as a class 2 Bluetooth device is used, and the sound quality is excellent. Figure 10(b) shows a UED NAM microphone used with an FM transmitter. The microphone output is fed into a one-chip FM-transmitter LSI device, and the listener receives the NAM signal through an ordinary FM radio. The transmission range is 30 m, and the RF signal output level is 0 dBm. The sound quality of the FM system is poorer than that of the Bluetooth one; nevertheless, the speech is sufficiently intelligible. Various types of FM transmitters for portable audio players as well as FM radios on the market with reasonable prices can be used.

Insert Fig.10 (a),(b) here

5.2 Type 3: transformation of silent-speech into whispered speech

The third type is an integrated system that uses numerical speech conversion. The weak vocal-tract resonance signals sensed by an SS NAM microphone are transformed into whispered speech using statistical conversion (Nakagiri *et al.* 2006, Nakamura *et al.* 2006, 2007).

There are several reasons we chose whispered speech as the output for this system. First is the common problem of using an electrolarynx — it is difficult to control the vibration frequency, either manually or automatically. Pitch plays an important role in speech as it carries prosodic features and information about the speaker's identity. Although a number of pitch control methods have been reported, the majority of EL devices do not permit intensity or pitch to be varied, so the speech they produce is monotonous. Second, synthesizing the natural pitch patterns needed to drive the EL is difficult. One could estimate the pitch pattern of a sentence, but only if the sentence were known before it was uttered. In addition, estimating the pitch pattern reflecting the speaker's emotional state would be particularly difficult. Consequently, developing a pitch control method for EL devices is still a challenge for future research.

In contrast, despite a lack of physical fundamental frequency (F0) components, whispered speech carries prosodic features and information about speaker's emotional state fairly well. This is because we are able to generate "pitch" for whispered speech in the brain, so to speak, with guidance from our implicit knowledge that the formant frequencies of vowels tend to correlate with F0 (Fant 1970, Titze 1989, Fitch and Giedd 1999, Higashikawa and Minifie 1999).

Figure 11 shows a block diagram of the speech transformation subsystem. Natural whisper sounds uttered by a male speaker were used as the target speech, and body-conducted speech generated by a small vibrator were used as the source speech. We used statistical speech conversion with a Gaussian mixture model (GMM) (Stylianou *et al.* 1998, Toda *et al.* 2005a, 2005b, 2007). In the training part, the joint probability density of the input and output features was modeled with a GMM using around 50 utterance pairs of input and output speech. In the conversion part, output features were determined using maximum likelihood estimation of the conditional probability density of output features for the given input features. Once we train the GMM, any sample of the input feature can be converted into that of the output feature.

Insert Fig.11 here

Figure 12 shows waveforms and spectrograms of source speech and transformed speech. The Japanese sentence spoken was "demo hoNtouni sounaNdesu"; the meaning is "but it is actually so". The transformed speech had a wider spectral range than the source speech and was free of pop noises. The mean mel-cepstral distortion between the transformed speech and the target whisper speech was 5 dB. The transformed voices were quite natural and intelligible compared with amplified weak vocal-tract resonance sounds.

Insert Fig.12 here

---

6. Evaluation of silent-speech enhancement systems

6.1 Type 1: simple amplification

We evaluated the first type system, simple amplification of the weak vocal-tract resonance signals, by testing its use by voice impaired and non-impaired volunteers. It was found useful even by the non-impaired volunteers when they had to talk at meetings or in classes while suffering hoarseness or a sore throat due to a cold. Both groups found speech articulation to be difficult without proper auditory feedback. This was expected because the produced voice itself is almost non-audible as the system uses a weak vibrator. In addition, voice impaired volunteers who had been using an EL found that it was hard to find the best spot to place the vibrator. They typically move the EL around their throats to find the best spot, and this "sweet spot" changes with time, requiring small adjustments to the position. With our system, such adjustments cannot be made without auditory feedback. The

amplified voice should have been fed back to the speaker via earphones with an adequate loudness level.

6.2 Type 2: simple amplification with wireless transmission

A married couple, an 82-year-old woman who could produce only a weak whisper as a result of a brain hemorrhage and her husband who had slight hearing loss, volunteered to use a prototype of the wireless system (Bluetooth Hi-Fi audio transmitter and Bluetooth headphones). They were able talk to each other comfortably at home with the system, thereby demonstrating the effectiveness of the wireless system. In this case, she did not request any additional auditory feedback, as she could hear her own whispered voices. As is often the case, simple devices excel at applicability: the combination of a NAM microphone with commercially available wireless transmitters/receivers is a simple and effective option for silent-speech enhancement.

6.3 Type 3: transformation of silent speech into whispered speech

Evaluation of the third type system was done by conducting two experiments.

In the first experiment, a non-impaired man who had spent 21 days learning how to speak with an EL, read 70 sentences from a newspaper in natural speech, in whispered speech, with an ordinary EL, and with the transformation system. The four types of speech datasets created were thus natural speech (NS), whispered speech (WS), electrolarynx speech (ELS), and transformed whispered speech from a weak body-conducted vocal-tract resonance signal driven by a 100-Hz pulse train using a small vibrator (TWS-SV-B). The GMM of the speech transformer was trained using 50 of the sentences in each dataset, and the remaining sentences were used for the listening test. Six non-impaired listeners subjectively evaluated the naturalness and intelligibility of the 20 test sentences for the four types of speech on a 5-point scale (1: bad – 5: excellent).

In the second experiment, a voice impaired man who had been using an EL for ten years read 100 ATR phoneme-balanced sentences with an ordinary EL and with a small vibrator. The four types of speech datasets created were ELS, transformed whispered speech from air-conducted electrolarynx speech (TWS-ELS-A), transformed whispered speech from body-conducted electrolarynx speech (TWS-ELS-B), and TWS-SV-B. The GMMs for the three ELS datasets were trained using 42 of the sentences in the dataset, and 7 sentences in the dataset were used for the listening test. The GMM for the SV dataset was trained using 90 of the sentences in the dataset, and the 10 remaining sentences were used for the listening test. In order to generate good transformed whispered speech, more data were necessary for training the GMM for the SV dataset than the ELS datasets. Three non-impaired listeners subjectively evaluated the naturalness and intelligibility of the test sentences for the four types of speech on a 5-point scale (1: bad – 5: excellent).

The mean opinion scores (MOS) for the two experiments are shown in Figures 13(a) and (b), respectively. In the first experiment, the transformed speech using a small vibrator was judged to be more natural and intelligible than the electrolarynx speech although not as natural and intelligible as the whispered speech. This result indicates that transformation into whispered speech from body-conducted weak vocal-tract resonance sound is a valid approach. In the second experiment, the electrolarynx speech was judged to have the highest intelligibility but the lowest naturalness. The transformed speech using an EL was judged to be more natural than the speech using a small vibrator although not as intelligible as the electrolarynx speech. Despite training data for the GMM is double the amount, intelligibility of TWS-SV-B is much worse than that of ELS. These results indicate that transformation into whispered speech does not necessarily improve intelligibility compared with using an EL alone.

Insert Fig.13 (a),(b) here

Evaluations for the transformed whispered speech of a non-impaired and a voice impaired man showed conflicting results. The EL proficiency of the speakers is one factor accounting for these conflicting results. The non-impaired volunteer was a novice EL user and thus was not a skillful user. In contrast, the voice impaired volunteer was an expert at using it and could thus produce highly intelligible speech with it. As he was fully accustomed to using an ordinary EL, he could have been

less attuned to using a small vibrator for producing speech. It should be noted that a proficient user can produce highly intelligible speech with an ordinary electrolarynx, and this is an example of the outstanding capability people have to adapt to new technologies. However, performance during the initial stage of adaptation is typically worse than with the conventional technology because the adapter is much more familiar with the older technology due to its longer period of use.

Another factor accounting for these conflicting results is the performance of the speech transformation model, which is trained so as to transform artificial speech into natural whispered speech. It is almost impossible to use the voice impaired speaker's own speech data to train the model unless the data are recorded prior to the loss of the speaker's voice. Although the GMM-based speech transformation algorithm can be expanded to handle completely different input and output speech datasets, more work is needed to improve the transformation performance.

7. Discussion

Usual voice that we hear or a microphone detects is air-conducted sound. We hear the sound via air conduction and a microphone detects the vibration of air by the sound. We also hear part of the sound via bone conduction. It is bone-conducted sound. Furthermore, there is body-conducted vocal-tract resonance sound, although it is very weak.

These diverse types of sounds can be classified using the two axes shown in Figure 14. One axis is the phonation type of the voice sound source. The major sound source is either quasi periodic glottal flow or random noise flow created by the air stream passing through the glottis from the lungs. Obviously, supraglottal sources also play an important role when obstruents are produced. The sound sources are then filtered by the vocal tract and radiated into the air from the mouth. Higher air pressure is needed for shouting or singing loudly, while lower air pressure is sufficient for whispering. Thus, this axis also represents speech energy. The other axis is the conduction medium of the voice sound. Typical air-conducted speech can be picked up with a regular microphone. Bone-conducted speech can be sensed with a bone-conduction microphone, such as a throat or ear microphone. Body-conducted speech can be sensed with a NAM microphone. The horizon of speech technology has expanded with this new type of silent speech.

Insert Fig.14 here

As described in Sections 2 and 3, vocal-tract resonance signals attenuate significantly in the body, and NAM microphones have bandpass characteristics. Therefore, higher frequency components are barely captured if there is a high SNR. NAM microphones, however, have fairly high sensitivity and broad frequency response as long as the target is telephone-band speech signals. That is, they function well as an input device for a silent communication system. Whether a wider bandwidth is necessary depends on the application. Broadening the frequency range of a NAM microphone, if necessary, as well as improving the SNR of the sensed NAM signals are tasks for future work.

Soft-silicone- and urethane-elastomer-type NAM microphones have considerable endurance. As these soft polymer materials are hydrophobic, the diaphragm of the ECM does not rust, and a NAM microphone generally works well for several years. Investigating the secular change in NAM microphones remains, however, for future work.

The NAM microphones developed so far use an ECM as a sensor unit because they are readily available at low cost. Since ECMs were originally designed for detecting small and fast vibration of the air, they are not necessarily the best device for detecting tissue vibration because of its higher acoustic impedance than air. Ceramic microphones or acceleration transducers are potential alternatives for NAM microphones. This alternative depends on whether they meet both the sensitivity and availability requirements. Searching for the best sensor unit for a NAM microphone also remains for future work.

8. Conclusion

We investigated the physical characteristics of weak body-conducted vocal-tract resonance signals called non-audible murmur (NAM) and the acoustic characteristics of two sensors we developed for

detecting these signals. The sensors have a sensitivity of between −41 and −58 dB [V/Pa] at 1 kHz, and the mean signal-to-noise ratio of the detected signals was 15 dB. On the basis of these investigations, we developed three types of silent-speech enhancement systems: (1) simple, direct amplification of weak vocal-tract resonance signals using a wired urethane-elastomer NAM microphone, (2) simple, direct amplification using a wireless urethane-elastomer-duplex NAM microphone, and (3) transformation of the weak vocal-tract resonance signals sensed by a wired soft-silicone NAM microphone into whispered speech using statistical conversion. Field testing of the systems showed that they enable voice impaired people to communicate verbally using body-conducted vocal-tract resonance signals. Listening tests demonstrated that weak body-conducted vocal-tract resonance sounds can be transformed into intelligible whispered speech sounds. Using these systems, people with voice impairments can re-acquire speech communication with less effort.

Acknowledgements
This work was supported by SCOPE of the Ministry of Internal Affairs and Communications of Japan.

Audio Files
The audio files for the utterances displayed in Figs. 6, 9 and 12 are available at http://auris.pu-toyama.ac.jp/ResearchProject_E.html.

References

Abe, M., Sagisaka, Y., Umeda, T., Kuwabara, H., 1990. Speech Database User's Manual, ATR Technical Report TR-I-0166, ATR Interpreting Telephony Research Lab.

Espy-Wilson C., Chari, V., Macauslan, J., Huang, C., and Walsh, M., 1998. Enhancement of electrolaryngeal speech by adaptive filtering, J. Speech. Lang. Hearing Res., 41(6), 1253-1264.

Fant, G. (1970). Acoustic Theory of Speech Production, 2nd ed. (Mouton, Paris).

Fitch, W., Giedd, J. (1999). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. 106, 1511-1522.

Fujisaka, Y., Nakagawa, S., Ogita, T., Tonoike, M., 2004. Analysis of wave propagation for bone-conducted ultrasonic in the heterogeneous human head model, Tech. Rep. IEICE, 103 (608), 13-17.

Heracleous, P., Nakajima, Y., Lee, A., Saruwatari, H., Shikano, K.., 2003. Accurate hidden Markov models for non-audible murmur (NAM) recognition based on iterative supervised adaptation, Proc. ASRU, 73-76.

Higashikawa, M., Minifie, F., 1999. Acoustical-Perceptual Correlates of "Whisper Pitch" in Synthetically Generated Vowels, Journal of Speech, Language, and Hearing Research 42, 583-591

Yoshinobu Kikuchi, Hideki Kasuya, 2004. Development and Evaluation of Pitch Adjustable Electrolarynx, Proc. International Conference on Speech Prosody, Nara, 761-764.
Nakajima, Y., Kashioka, H., Shikano, K., Campbell, N., 2003a. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin, Proc. ICASSP, 708-711.

Nakajima, Y., Kashioka H., Shikano, K., and Campbell, N., 2003b. Non-Audible Murmur Recognition, Proc. EUROSPEECH, 2601-2604.

Nakajima, Y., 2005. Development and evaluation of soft silicone NAM microphone, Tech. Rep. IEICE, 105 (97), 7-12.

Nakajima, Y., Kashioka H., Shikano, K., and Campbell, N., 2005. Remodeling of the Sensor for Non-Audible Murmur (NAM), Proc. INTERSPEECH, 389-392.

Nakajima, Y., Kashioka, H., Campbell, N., Shikano, K., 2006. Non-audible murmur (NAM) recognition, IEICE Trans. on Information and System, E89-D (1), 1-8.

Nakagiri, M., Toda, T., Kashioka, H., Shikano, K., 2006. Improving Body Transmitted Unvoiced Speech with Statistical Voice Conversion, Proc. INTERSPEECH - ICSLP, 2270-2273.

Nakamura, K., Toda, T., Saruwatari, H., Shikano, K., 2006. Speaking Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech, Proc. INTERSPEECH - ICSLP, 1395-1398.

Nakamura, K., Toda, T., Saruwatari, H., Shikano, K., 2007. Impact of various small sound source signals on voice conversion accuracy in speech communication aid for Laryngectomees, Proc. INTERSPEECH- EUROSPEECH, 2517-2520.

Nota, Y., Kitamura, T., Honda, K., Takemoto, H., Hirata, H., Shimada, Y., Fujimoto, I., Shakudo, Y., Masaki, S., 2007. A bone-conduction system for auditory stimulation in MRI, Acoust. Sci. & Tech., 28, 33.38.

Oestreicher, H.L., 1951. Field and Impedance of an Oscillating Sphere in a Viscoelastic Medium with an Application to Biophysics, J. Acoust. Soc. Am., 23(6), 707-714.

Otani, M., Hirahara, T., Shimizu, S., Adachi, S., 2009. Numerical simulation of transfer and attenuation characteristics of soft-tissue conducted sound originating from vocal tract, Applied Acoustics 70, 469-472.

Sagisaka, Y., Uratani, N., 1992. ATR spoken language database, J. Acoust. Soc. Japan 48, 878-882.

Stylianou, Y., Cappe, O., Moulines, E., 1998. Continuous probabilistic transform for voice conversion, IEEE Trans. SAP, 6 (2), 131-142,

Titze, I. (1989). Physiologic and acoustic differences between male and female voices, J. Acoust. Soc. Am. 85, 1699-1707.

Toda, T., Black, A., Tokuda, K., 2005a. Spectral Conversion Based on Maximum Likelihood Estimation Considering Global Variance of Converted Parameter, Proc. ICASSP, 1, 9-12.

Toda, T., Shikano, K., 2005b. NAM-to-speech conversion with Gaussian mixture models. Proc. INTERSPEECH, 1957-1960.

Toda, T., Black, A., Tokuda, K., 2007. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory, IEEE Trans. ASLP, 15 (8), 2222-2235.

Uemi, N. , Ifukube, T., Takahashi, M., Matsushima, J., 1994. Design of a new electrolarynx having a pitch control function, Proc. IEEE International Workshop on Robot and Human Communication, Nagoya, 198-203.

Figure captions

Figure 1: (a) Geometry of simulated region including head, sound source, and receivers. (b) Vocal-tract transfer functions I_{in}/I_{out} and I_{nam}/I_{out} correspond to sound intensity level observed at NAM microphone assuming sound radiated from mouth has flat frequency characteristics.

Figure 2: Photographs and structures of three types of NAM microphones examined: (a) soft-silicone (SS), (b) urethane-elastomer (UE), and (c) urethane-elastomer-duplex (UED).

Figure 3: X-ray computed tomographic images of (a) SS NAM microphone and (b) UED NAM microphone.

Figure 4: Block diagram of system for calibrating NAM microphones.

Figure 5: Mean frequency response of NAM microphones.

Figure 6: Waveform and spectrogram of NAM signals for (a) male and (b) female speaker recorded with SS NAM microphone.

Figure 7: Long-term average spectra (LTAS) of NAM signals resampled at 16 kHz.

Figure 8 Long-term average spectra (LTAS) of underlying NAM signals were calculated by subtracting frequency responses of NAM microphone and microphone amplifier from LTAS shown in Figure 7.

Figure 9: Waveforms and spectrograms of NAM signals recorded simultaneously using (a) SS and (b) UED NAM microphones.

Figure 10: Photographs of wireless NAM microphones: (a) Bluetooth system and (b) FM transmitter system.

Figure 11: Block diagram of voice transformation subsystem based on GMM.

Figure 12: Waveform and spectrogram of (a) source NAM signal driven by a small vibrator and (b) speech transformed into whispered voice.

Figure 13: Mean opinion score of transformed speech: (a) non-impaired speaker who was novice electrolarynx (EL) user and (b) voice impaired speaker who was skillful EL user.

Figure 14: Classification of diverse types of speech.

Table 1: Specifications of three NAM microphones

Figure 1

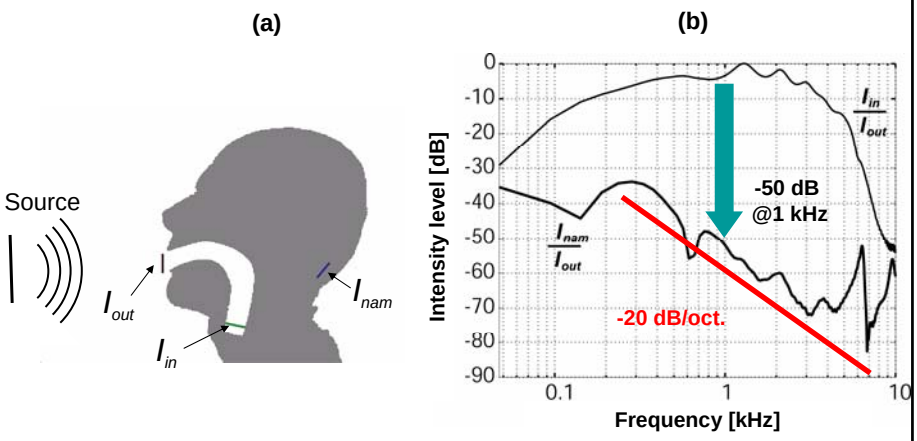

Figure 2

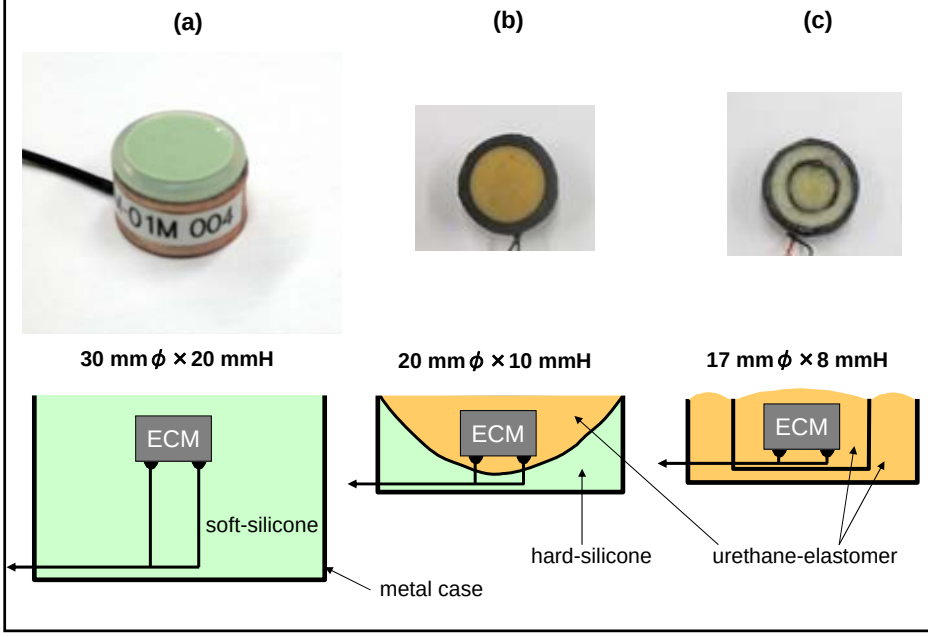

Figure 3

X-ray computed tomography scanning images of (a) a soft-silicone-type NAM microphone (SS), and (b) a urethane-elastomer-duplex-type NAM microphone (DUE).

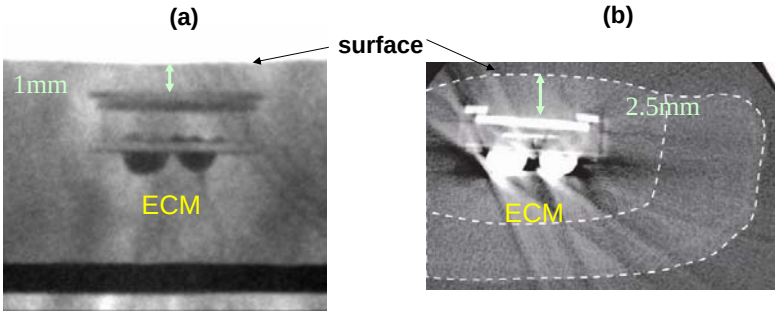

Figure 4

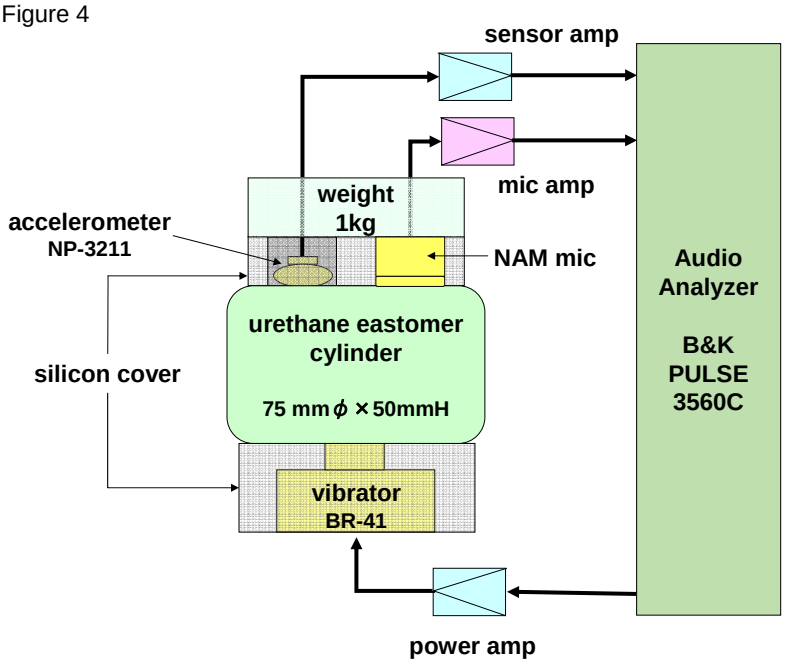

Figure 5

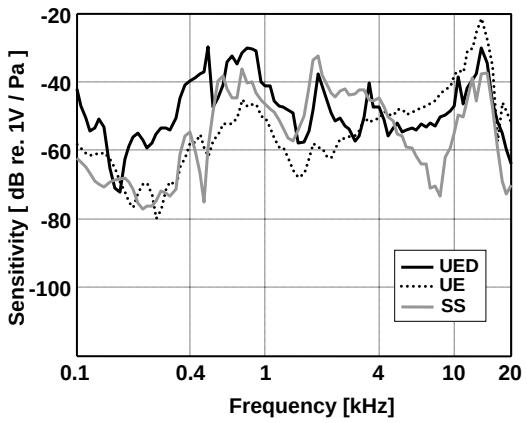

Figure 6

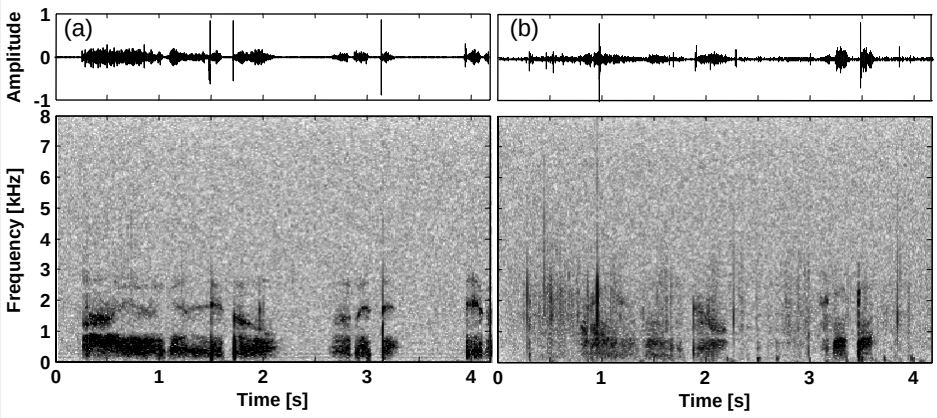

Figure 7

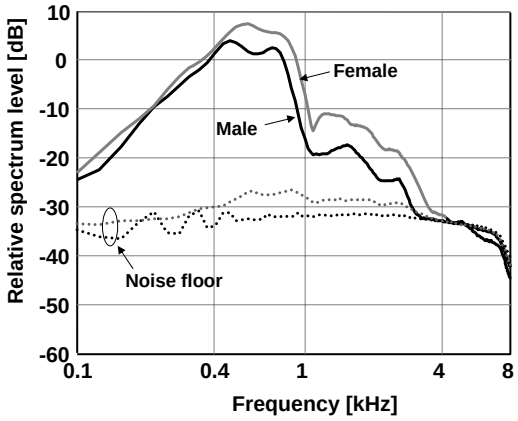

Figure 8

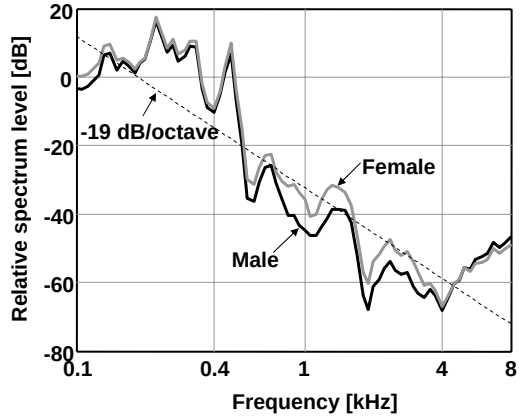

Figure 9

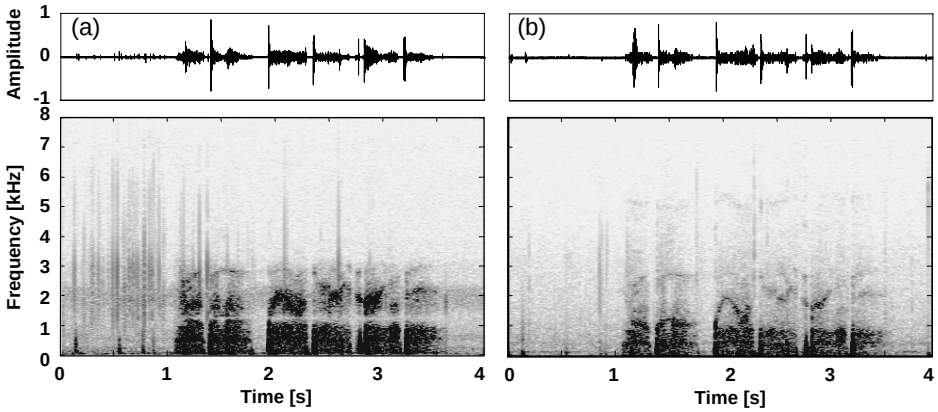

Figure 10

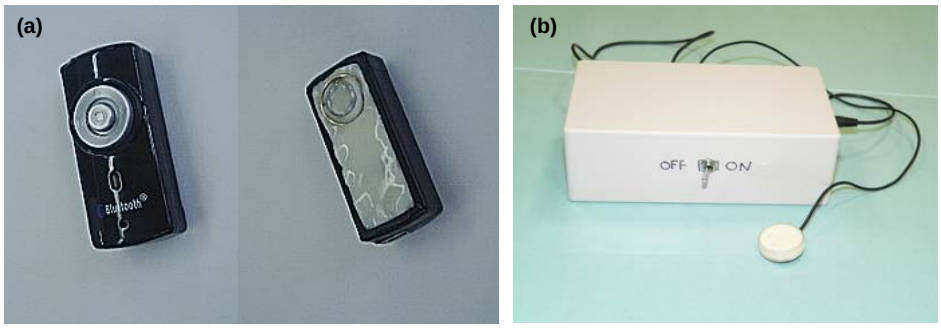

Figure 11

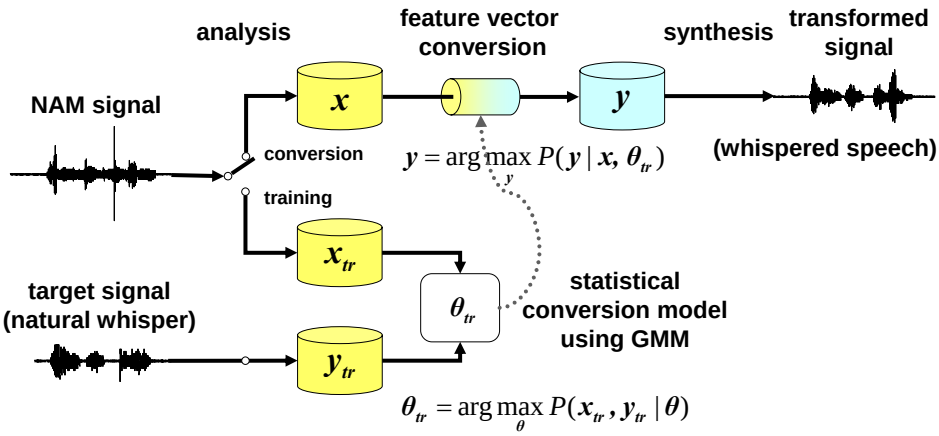

Figure 12

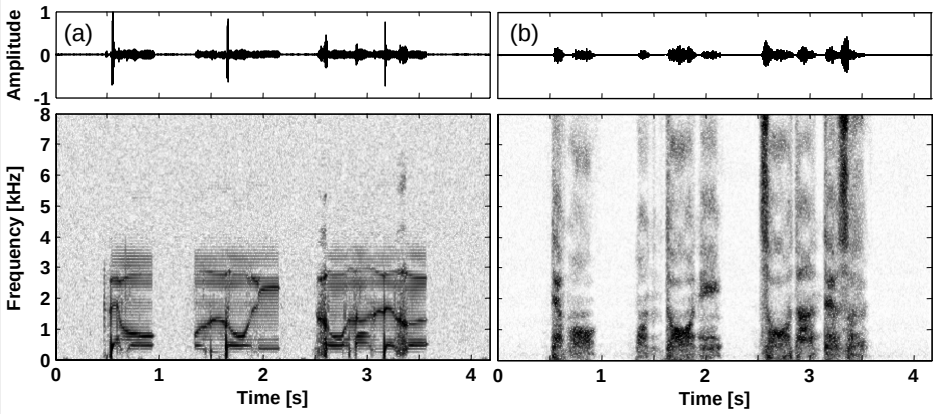

Figure 13

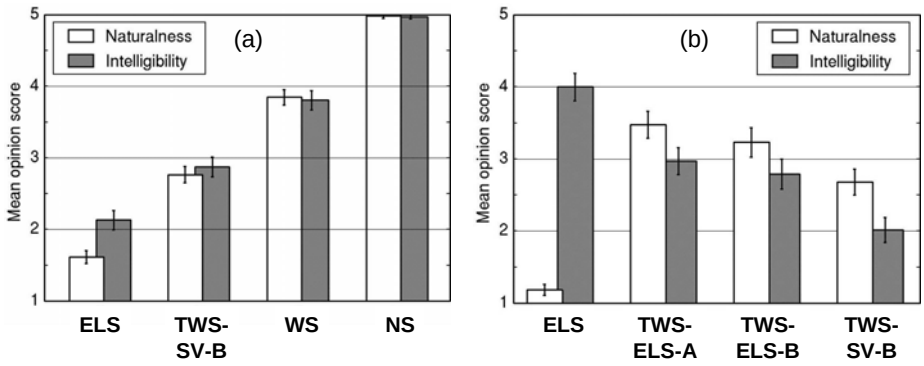

Figure 14