INTERSPEECH 2009 BRIGHTON

Technologies for Processing Body-Conducted Speech Detected with Non-Audible Murmur Microphone

Tomoki Toda, Keigo Nakamura, Takayuki Nagai,
Tomomi Kaino, Yoshitaka Nakajima, Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology, Japan
tomoki@is.naist.jp, kei-naka@is.naist.jp, shikano@is.naist.jp

Abstract

In this paper, we review our recent research on technologies for processing body-conducted speech detected with Non-Audible Murmur (NAM) microphone. NAM microphone enables us to detect various types of body-conducted speech such as extremely soft whisper, normal speech, and so on. Moreover, it is robust against external noise due to its noise-proof structure. To make speech communication more universal by effectively using these properties of NAM microphone, we have so far developed two main technologies: one is body-conducted speech conversion for human-to-human speech communication; and the other is body-conducted speech recognition for man-machine speech communication. This paper gives an overview of these technologies and presents our new attempts to investigate the effectiveness of body-conducted speech recognition.

Index Terms: silent speech, Non-Audible Murmur, body-conducted speech, voice conversion, automatic speech recognition

10.21437/Interspeech.2009-224

1. Introduction

In recent decades the style of speech communication has dramatically changed due to the development of information technologies: e.g., the explosive spread of cell phones has enabled people to talk with each other beyond limitations of distance and space. Those technologies have brought more convenient means of speech communication to us.

Can we really communicate with speech any time? There are actually some situations where we face difficulties with speech communication. For instance, we have trouble privately talking in the crowd; speaking itself sometimes annoys others in quiet environments; we may lose our voices if given surgery to remove speech organs such as larynx due to laryngeal cancer. Many barriers still exist in speech communication. The development of technologies to overcome these inherent problems of speech communication is essential to make our speech communication more universal.

Recently *silent speech interfaces* have attracted attention as a technology to support new speech communication styles. They enable speech communication to take place without the necessity of emitting an audible acoustic signal. There have been several attempts to explore sensing devices alternative to air microphone, such as a throat microphone [1], electromyography (EMG) [2], and ultrasound imaging [3]. These sensing devices are effective for soft speech in a private talk and as a speaking aid for the vocally handicapped. In addition, they are also effective for noise robust speech communication as Subramanya *et al*. [4] have reported that bone-conducted speech signals are very informative to enhance speech sounds under heavy noise conditions. These devices will bring a new paradigm to speech communication.

As one of the microphones to detect body-conducted speech, Non-Audible Murmur (NAM) microphone has been developed by Nakajima *et al*. [5]. Inspired by a stethoscope, NAM microphone was originally developed to detect extremely soft murmur called NAM, which is so quiet that people around the speaker hardly hear it. Placed on the neck below the ear, NAM microphone detects various types of speech such as NAM, whisper, and normal speech through the soft tissue of the head. It is robust against external noise due to its noise-proof structure like the other body-conductive microphones. Moreover, its usability is better compared with other devices such as EMG or ultrasound systems. Considering these properties, we focus on NAM microphone as one of the promising devices.

NAM microphone enables us to talk in various types of body-conducted speech according to situations, e.g., NAM for *silent speech communication* or body-conducted normal speech for noise robust speech communication. However, there are some serious problems of using NAM microphone in speech communication. One of the biggest problems is the severe degradation of speech quality caused by essential mechanisms of body conduction such as lack of radiation characteristics from lips and influence of low-pass characteristics of the soft tissue. Therefore, quality improvements of body-conducted speech are essential to use it as a human-to-human speech communication medium. To deal with this problem, we have proposed statistical approaches to body-conducted speech enhancement [6].

Body-conducted speech detected with NAM microphone is also useful in man-machine speech communication. External noise is always problematic for the speech interface. NAM microphone dramatically alleviates this problem. Moreover, it also works as a silent speech interface allowing us to quietly input words to a machine. These properties of NAM microphone would make it possible to develop a more universal speech interface dealing with a wide variety of our speaking styles used as the situation demands. For this purpose, we have developed a body-conducted speech recognition system by building acoustic models for body-conducted speech.

In this paper, we review our recent research on development of technologies for processing body-conducted speech detected with NAM microphone. We also present our new attempts to investigate the effectiveness of body-conducted speech recognition.

2. Non-Audible Murmur (NAM) microphones

NAM is defined as the articulated production of respiratory sounds without vibration of vocal folds, which can be transmitted through only the soft tissue of the head [5]. It is hardly audible because its power is extremely low. Such an extremely soft speech signal is relatively difficult to be detected with a

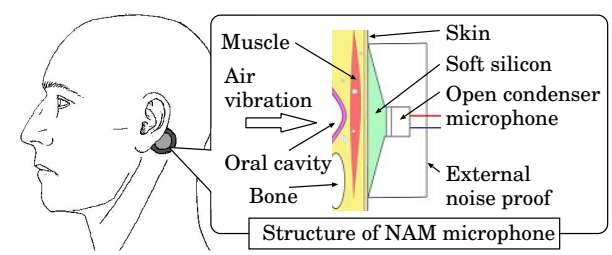

Figure 1: *Setting position and structure of NAM microphone.*

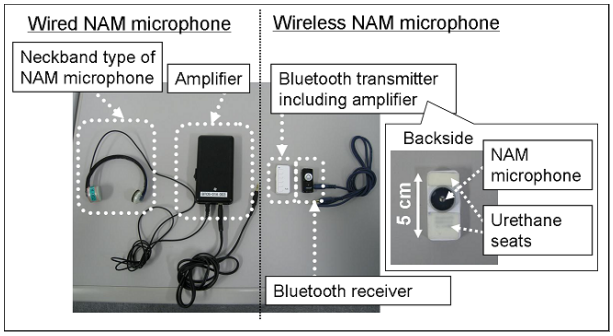

Figure 2: *Wired- and Wireless-types of NAM microphone.*

usual air-conductive microphone because it is easily vanished into external noise.

NAM microphone has been designed to detect high-quality NAM [5]. It is attached directly to the talker's body. **Figure 1** shows the best position to attach the microphone. From this position, air vibrations in the vocal tract are captured through only the soft tissue of the head. This position enables recording of extremely soft voices with good quality by evading the transmission through obstructions such as a bone whose acoustic impedance is quite different from that of the soft tissue. Moreover, NAM microphone has a special structure as shown in **Fig. 1**. Soft silicone, whose acoustic impedance is close to that of the soft tissue, is used as the medium between body and a condenser microphone for alleviating loss of conduction [7]. Furthermore, a noise-proof cover effectively increases the signal-to-noise ratio of body-conducted speech under noisy conditions.

Database building is essential for developing technologies for processing body-conducted speech. To record a large amount of body-conducted speech with consistent quality, we need to develop several NAM microphones of which characteristics are stable enough. We have so far made two typical prototypes in cooperation with a few Japanese companies. One is a wired-type and the other is a wireless-type as shown in **Fig. 2**. They enable us to record body-conducted speech as consistently as possible.

3. Body-Conducted Speech Conversion

Statistical voice conversion, which has originally been proposed for speaker conversion, is one of useful techniques to enhance the body-conducted speech. This technique converts voice characteristics of input speech into those of some other speech while keeping linguistic information unchanged. A conversion model capturing correlations between acoustic features of source and target voices is trained in advance using a small amount of parallel data consisting of utterance pairs of these two voices. The trained model allows the conversion from any sample of the source into that of the target using only acoustic information. It is well known that a Gaussian mixture model (GMM) of the joint probability density of the source and target works reasonably well as the conversion model [8].

We have proposed several conversion methods for enhancing various types of body-conducted speech [6] based on a state-of-the-art GMM-based conversion technique, which allows probabilistic conversion considering both the inter-frame correlation and the higher-order moment (details in [9]). In this approach, it is essential to select an appropriate target speech style according to the speaking style of body-conducted speech.

For *silent speech communication*, conversion methods of enhancing body-conducted unvoiced speech [10, 11] have been proposed. The first attempt is to convert NAM into normal speech [10]. Not only spectral features of normal speech but also its excitation features such as F_0 are estimated from only spectral features of NAM. This method generates the converted speech of which voice quality is similar to that of the target natural speech. However, the main weakness of this method is difficulties of the F_0 estimation from spectral features of unvoiced speech. To avoid this problem, we have further proposed the conversion method into whisper that is familiar unvoiced speech [11]. This method yields significant improvements in naturalness and intelligibility compared with the original NAM.

For noise robust speech communication, we have proposed conversion methods of enhancing body-conducted voiced speech [6]. Quality of body-conducted normal speech is significantly improved if converted into normal speech. In this conversion, F_0 values and unvoiced/voiced information extracted from body-conducted normal speech are accurate enough to be directly used in synthesizing the converted normal speech. We have also proposed conversion from a body-conducted soft voice into normal speech for supporting private talk in public areas. Interestingly this conversion doesn't cause any significant quality improvements. Several voiced phonemes are often devoiced in a soft voice. Therefore, there are noticeable unvoiced/voiced mismatches between a soft voice and normal speech, and they are not straightforwardly compensated. This problem is effectively avoided by using the soft voice as the conversion target. Consequently, quality of a body-conducted soft voice is significantly improved by the conversion into a soft voice.

The body-conducted speech conversion technique is also effective for speaking aid. Problems of an existing electrolarynx for laryngectomees are a leakage of loud external excitation signals and mechanical sounds of the generated voices. To address these problems, we have proposed a new speaking aid system for laryngectomees based on three main technologies: 1) generation of external excitation signals with extremely small power; 2) detection of body-conducted artificial speech with NAM microphone; and 3) voice conversion into whisper [12]. This system enables laryngectomees to speak in whisper, which sounds more natural than the artificial voice generated by the electrolarynx, while keeping emitted excitation signals less audible.

4. Body-Conducted Speech Recognition

The main difference between a normal speech recognition system and a body-conducted speech recognition system is acoustic models. Therefore, we need to build specific acoustic models for body-conducted speech. Conventional adaptation techniques such as Maximum Likelihood Linear Regression (MLLR) [13] work reasonably well for developing hidden Markov models (HMMs) for NAM from those for normal speech. It has been reported that iterative MLLR adaptation process using the adapted model as the initial model at the next

EM-iteration step is very effective because acoustic characteristics of NAM are considerably different from those of normal speech [14].

We have previously demonstrated the performance of automatic NAM recognition for only a few specific speakers. In fact they are very special because they have learnt how to utter NAM so as to be well recognized by the recognition system through their own research experiences on NAM recognition. Therefore, it is still questionable how much recognition accuracy is obtained for general speakers. Moreover, only two types of body-conducted speech (i.e., NAM and body-conducted normal speech) have been coped with in our previous work. As mentioned above, one important feature of NAM microphone is the capability of detecting a wide variety of body-conducted speech. Therefore, it is worthwhile to investigate recognition performance of body-conducted speech for various speakers and various speaking styles.

4.1. NAM Recognition for General Speakers

In this paper, we investigate NAM recognition performance for general speakers who are not familiar with NAM. We have used the speaker-independent model for normal speech (*SI-Normal*) as the initial model in our previous work. It is well known that recognition performance of the adapted model is affected by the initial model. Therefore, we exploit NAM data uttered by many general speakers effectively for building the better initial model. Two standard approaches are investigated: 1) speaker-independent NAM model (*SI-NAM*) trained with those NAM data; and 2) a canonical model for NAM adaptation (*SAT-SI-Normal*) trained using those NAM data in speaker adaptive training (SAT) paradigm [15]. *SI-Normal* is used as the initial model in both approaches.

4.2. Recognition in Various Speaking Styles

In this paper, we investigate recognition performance of body-conducted speech uttered in multiple speaking styles including NAM, whisper, a soft voice, and normal speech. To flexibly recognize them, we adopt two approaches: 1) a style-mixed model trained using data of all styles simultaneously, and 2) parallel decoding with the individual style-dependent models separately trained for individual styles, in which a recognition result of the model with the highest likelihood for the input utterance is automatically selected.

Considering the use of body-conducted speech interfaces in practical situations, it is also essential to investigate its performance under noisy conditions. It has been reported that the Lombard reflex causes severe degradation of NAM recognition performance [16]. As an initial attempt to cope with this problem, we apply the above two approaches to body-conducted speech recognition under noisy conditions regarding body-conducted speech uttered under different noise levels as different speaking styles. Only two types of body-conducted speech, i.e., voiced and unvoiced, are considered under each noise level because it is hard to distinctively speak in each of NAM and whisper or each of a soft voice and normal speech under noisy conditions.

5. Experimental Evaluations of Body-Conducted Speech Recognition

We conducted large vocabulary continuous speech recognition experiments to evaluate 1) NAM recognition performance for general speakers, and 2) recognition performance of body-conducted speech in various speaking styles.

5.1. Evaluation of NAM Recognition for General Speakers

We recorded NAM data from 58 general speakers. During their recording, we briefly told each speaker how to utter NAM and checked if each speaker uttered in NAM properly.

We adopted 12 MFCCs, 12 ΔMFCCs and Δ power as the acoustic features. Left-to-right 3 state triphone HMMs with no skip were used as acoustic models. The number of shared states was 2189 and the state output probability distribution was modeled with 16 mixture components of GMMs. We used 60 k word trigram language model trained with Japanese newspaper articles.

Twenty NAM utterances were selected from Japanese newspaper articles as a test set for each speaker so that perplexity and out of vocabulary words in each test set were as constant as possible over different speakers. All of remaining NAM utterances (about 130 to 220 utterances per speaker) were used as the training or the adaptation data. We conducted 6-fold cross-validation test using all 58 speakers' data.

SI-Normal was built with normal speech database designed for training speaker-independent model, which included voices of several hundreds of speakers. *SI-NAM* and *SAT-SI-Normal* were built with all speakers' NAM data included in each cross-validation training set. Finally, the speaker-dependent models for individual speakers in each cross-validation test set were built from these three initial models using iterative MLLR mean and variance adaptation.

Figure 3 shows the relationship of word accuracy of the speaker-dependent models for individual speakers between when using the conventional initial model (*SI-Normal*) and when using the new initial models (*SI-NAM* and *SAT-SI-Normal*). Word accuracy averaged over all speakers and its standard deviation are 64.43$\pm$14.81% for *SI-Normal*, 67.61$\pm$12.09% for *SI-NAM*, and 72.58$\pm$11.24% for *SAT-SI-Normal*, respectively. We can see that better speaker-dependent models are obtained by using NAM data of many other speakers for training the initial model. We can also see that word accuracy varies widely over different speakers. Although *SI-NAM* sometimes gives rise to worse speaker-dependent models compared with *SI-Normal*, *SAT-SI-Normal* yields better ones more consistently. Consequently, the inter-speaker variation of recognition performance is significantly reduced by SAT. However, the reduced variation is still large. The further reduction would be essential in the development of the NAM recognition interface.

5.2. Evaluation of Body-Conducted Speech Recognition in Various Speaking Styles

We used body-conducted speech data uttered by one female speaker in four speaking styles, NAM, whisper, a soft voice, and normal speech, under clean conditions (in a sound-proof room). In addition, we used body-conducted voiced or unvoiced speech uttered by the same speaker under noisy conditions. These data were recorded by presenting each of 50 dBA, 60 dBA, and 70 dBA office noise to the speaker with a headphone. As a result, six types of data were recorded under noisy conditions.

Left-to-right 3-state phonetic tied mixture HMMs [17] with no skip were used. The number of tied mixture components was set to 64. The number of shared states was 3000. The vocabulary size of the trigram language model was 20 k.

We first conducted an experimental evaluation under clean

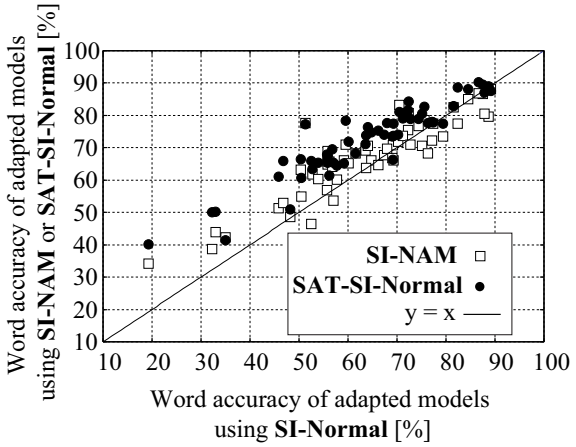

Figure 3: *Relationship of word accuracy of speaker-dependent models for individual speakers between when using 'SI-Normal'and when using 'SI-NAM' and 'SAT-SI-Normal'.*

Table 1: *Word accuracy for each speaking style when using matched style-dependent models 'Mathced', style-mixed model 'Mixed' and parallel decoding 'Parallel'.*

Clean conditions	Normal	Soft	Whisper	NAM
Matched	89.41	84.18	86.67	77.90
Mixed	87.40	84.74	81.04	75.80
Parallel	89.41	84.18	86.67	77.90

Noisy conditions	Voiced speech			Unvoiced speech		
Noie [dBA]	50	60	70	50	60	70
Matched	88.22	87.82	89.01	81.84	67.38	73.21
Mixed	86.21	85.54	86.35	77.20	60.88	62.35
Parallel	88.22	87.55	88.61	81.84	67.81	72.94

conditions using only data of the four speaking styles in clean conditions for model training. In these evaluations, we built a style-mixed model covering all of these four speaking styles and four style-dependent models for the individual speaking styles, which were used in the parallel decoding. And then we conducted another experimental evaluation under noisy conditions additionally using the six types of data in noisy conditions for model training. In these evaluations, we built a style-mixed model covering all of both the six types of data in noisy conditions and the four types of data in clean conditions. Ten style-dependent models including additionally trained six style-dependent models in noisy conditions were used for the parallel decoding. The iterative MLLR mean and variance adaptation was used to build the above body-conducted speech models from the speaker-independent normal speech model. We used 100 utterances as an adaptation set and 50 utterances as a test set for each style.

Table 1 shows the results. The style-mixed model tends to cause performance degradation compared with the matched style-dependent models especially in body-conducted unvoiced speech under noisy conditions. We also tried increasing the number of tied mixture components but this degradation was still observable. Results of the parallel decoding are very close to those of the matched style-dependent models because the selected model almost completely corresponds to the actual style of an input utterance. Interestingly results of 60 dBA are worse than the others. It is expected that such a noise level tends to make us speak more unsteadily compared with under quieter or louder conditions.

6. Conclusions

We have reviewed our recent research on development of technologies for processing body-conducted speech detected by Non-Audible Murmur (NAM) microphone: i.e., development of NAM microphone; body-conducted speech conversion; and body-conducted speech recognition. Moreover, we have further investigated the effectiveness of body-conducted speech recognition in various conditions.

Acknowledgment: This research was supported in part by MIC SCOPE and MEXT Grant-in-Aid for Scientific Research (A).

7. References

[1] S-C. Jou, T. Schultz, and A. Waibel. Adaptation for soft whisper recognition using a throat microphone. *Proc. INTERSPEECH*, pp. 1493–1496, Jeju Island, Korea, 2004.

[2] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel. Session independent non-audible speech recognition using surface electromyography. *Proc. ASRU*, pp. 331–336, San Juan, Puerto Rico, Nov. 2005.

[3] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, M. Stone. Continuous-speech phone recognition from ultrasound and optical images of the tongue and lips. *Proc. Interspeech*, pp. 658–661, Antwerp, Belgium, Aug. 2007.

[4] A. Subramanya, Z. Zhang, Z. Liu, A. Acero. Multisensory processing for speech enhancement and magnitude-normalized spectra for speech modeling. *Speech Communication*, Vol. 50, No. 3, pp. 228–243, 2008.

[5] Y. Nakajima, H. Kashioka, N. Cambell, and K. Shikano. Non-Audible Murmur (NAM) Recognition. *IEICE Trans. Information and Systems*, Vol. E89-D, No. 1, pp. 1–8, 2006.

[6] T. Toda, K. Nakamura, H. Sekimoto, K. Shikano. Voice conversion for various types of body transmitted speech. *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009.

[7] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. Remodeling of the sensor for non-audible murmur (NAM). *Proc. INTERSPEECH*, pp. 389–392, Lisbon, Portugal, Sep. 2005.

[8] Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142, 1998.

[9] T. Toda, A.W. Black, and K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Trans. Audio, Speech and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[10] T. Toda and K. Shikano. NAM-to-speech conversion with Gaussian mixture models. *Proc. INTERSPEECH*, pp. 1957–1960, Lisbon, Portugal, Sep. 2005.

[11] M. Nakagiri, T. Toda, H. Saruwatari, and K. Shikano. Improving body transmitted unvoiced speech with statistical voice conversion. *Proc. INTERSPEECH*, pp. 2270–2273, Pittsburgh, USA, Sep. 2006.

[12] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano. Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech. *Proc. INTERSPEECH*, pp. 1395–1398, Pittsburgh, USA, Sep. 2006.

[13] M.J.F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, Vol. 12, No. 2, pp. 75–98, 1998.

[14] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano. Accurate hidden Markov models for Non-Audible Murmur (NAM) recognition based on iterative supervised adaptation. *Proc. ASRU*, pp. 73–76, St. Thomas, USA, Dec. 2003.

[15] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. *Proc. ICSLP*, pp. 1137–1140, Philadelphia, Oct. 1996.

[16] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano. Investigating the role of the Lombard reflex in Non-Audible Murmur (NAM) recognition. *Proc. INTERSPEECH*, pp. 2649–2652, Lisbon, Portugal, Sep. 2005.

[17] A. Lee, T. Kawahara, K. Takeda, and K. Shikano. A new Phone Tied-Mixture model for efficient decoding. *Proc. ICASSP*, pp. 1269–1272, Istanbul, Turkey, June 2000.