NON-AUDIBLE MURMUR RECOGNITION

Yoshitaka Nakajima, Hideki Kashioka, Kiyohiro Shikano, Nick Campbell

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101 Japan

ABSTRACT

We propose a new style of practical input interface for the recognition of non-audible murmur (NAM), i.e., for the recognition of inaudible speech produced without vibration of the vocal folds. We developed a microphone attachment, which adheres to the skin, applying the principle of a medical stethoscope, found the ideal position for sampling flesh-conducted NAM sound vibration and retrained an acoustic model with NAM samples. Then using the Julius Japanese Dictation Toolkit, we tested the possibilities for practical use of this method in place of an external microphone for analyzing air-conducted voice sound. Additionally we propose laryngeal elevation index (LEI), a new index of prosody, which can show the prosody of NAM without F0, using simple processing of images from medical ultrasonography. We realized and defined NAM never used for input or communication and propose that we should make use of it for the interface of human-human and human-cybernetic machines.

1. INTRODUCTION

Anticipating the near future of information technologies, very few people would deny the following three trends: Further infiltration of wireless and broadband networking systems throughout the world, introduction of domestic robots of various types into our daily lives, and development of personal information processing terminals that are miniaturized and wearable.

Few people predicted the abrupt spread of cellular phones. Given that cellular phones are no less than small computers, each with a CPU, we presume that it will not be long before we individually use small personal digital assistants to communicate with other persons and with cybernetic machines through a global wired/wireless network. If this becomes reality, then what will be the best input interface under such circumstances? We are not really satisfied with the limited keys of a cellular phone that have already become most widely-used input device in the history of mankind. Speech recognition had been anticipated as the most natural input-interface for the future by a lot of people, both in the real world and in the worlds of science fiction, from well before the last century.
Speech recognition is now beginning to be more widely used, but only in very restricted domains such as car navigation systems or ticket vending machines. Nevertheless it is an accomplished and practicable art with a technological accumulation of about 30 years, so why don't we see people using speech recognition interfaces all around us?

2. WHY DON'T WE PUT SPEECH RECOGNITION INTO PRACTICAL EVERYDAY USE?

First of all, the major premise of speech recognition is that we can detect speech sounds from the air, using an external microphone in combination with digital sampling and processing. Therefore the technology is essentially prone to external noise. This becomes a serious problem when speech recognition is to be used outdoors, in public places, or inside vehicles. Secondly, and from an opposite standpoint, speech is often considered to be an unwanted noise source in the same public situations, especially in a busy office.

Naturally if the voice is to be audible, surrounding people can overhear the content of speech and thereby get to know what is being input to the recognizer. Because the sound of the voice is conducted through the air, we can't control its dispersal. Furthermore, can voice commands really be trusted in an emergency? The more urgent a situation becomes, the noisier the place usually gets, for example often with screaming and crying. Additionally, any person who has actually used speech recognition in public understands that it can be embarrassing to talk to a machine in the presence of others.

In quiet indoor desktop environments, the current applications of speech recognition on the market have a recognition accuracy that is good enough for practical use. But on the other hand, they can be difficult to use in the poor acoustic environment of domestic housing conditions such as we find in Japan. Once started with that wonderful hands-free interface, the user might soon have to give up listening to relaxing background music, or fluttering with the sudden entrance of family members because of the noise and embarrassment.

3. PROPOSAL FOR A NEW SENSING AND ANALYSIS METHOD

When the speech recognition technology was in its early development stages, the original developers probably did not think much about its use in the context of the above-mentioned three key technologies of wireless broadband networking, robots and handy wearable terminals. When thinking about speech recognition, they probably assumed that people would talk to robots and computers directly, as with regular human speech. Of course many people still imagine so. But now we can see that the essence of the problem restricting the spread of speech recognition is that we have to continuously detect voice sounds, which are dispersed in the air, by use of external microphones,

and have to rely on the vibrations carried in the air as the prime medium of voice transmission.

In contrast, we propose below a method by which we can recognize spoken input to a personal terminal at hand or 'on the skin', and then send the recognized parameters as texts to the terminals of others through the global network. However, we should continue to use the spoken language culture which people have learnt to rely on. Paradoxically speaking, the problem of speech recognition is in phonating speech. Perhaps we should no longer rely so much on talking to external microphones?

4. NON-AUDIBLE MURMUR RECOGNITION

Speech is one of the actions that originate inside the human body. One of the best methods of examining what is happening in the human body is to touch it, as medical doctors have learnt always to do first. We propose that it might be more efficient to analyze the original vibrations uttered as human speech from inside the body, through the surface of the skin, instead of analyzing the sounds in the air, after being discharged through the mouth. If this is possible, then we can put non-audible murmur into use as a new speech communication interface.

4.1. Non-audible murmur (NAM)

Non-audible murmur (NAM) is a term used to refer to the kind of speech action taking place in the mouth, which a nearby person would not be able to hear. In this study, it is defined as the articulated production of respiratory sound without recourse to vocal-fold vibration, produced by the motions and interactions of speech organs such as the tongue, palate, lips etc., which can be transmitted through the soft tissue of the head. Although it is difficult to define precisely the acoustic differences between "whisper" and "NAM", the term "whisper" implies that limited nearby listeners can hear the content of the speech, and that it can be recorded by an external microphone through transmission in the air, as reported in previous studies [1]. It becomes easier to understand NAM when we think of it as a whisper too small to be heard by anyone, rather like a personal silent prayer or a very quiet breathy voice, arising from the gestures within the mouth. We can say that whispers are the small part of highly powered NAMs (with hard narrowing of vocal tract), which are used for confidential communication to limited persons.

4.2. Basic concept of the proposed method

Figure 1 illustrates the basic concept of this method, whereby NAM is detected instead of oral speech for the speech recognition input interface. The NAM is detected using a stethoscopic microphone bonded to the surface of the skin, close behind the ear, and sent through the microphone amplifier by cable or wireless system to be sampled by the AD converter for digital processing. Parameters are extracted for recognition as they would be for a normal voiced speech source. We then analyze the parameters using a Hidden Markov Model (HMM), which is probably the most useful and accurate analysis method for real-time speech recognition up to the present time. Both the language model and the acoustic model are usually used cooperatively. But in this method we need to train a new NAM acoustic model, built up from multiple samples of NAM, in order to replace the existing speech acoustic models for accurate recognition. This is one of the most important points of our study. Once this operation is performed, we can apply almost all aspects of speech recognition technology that have been accumulated to date.

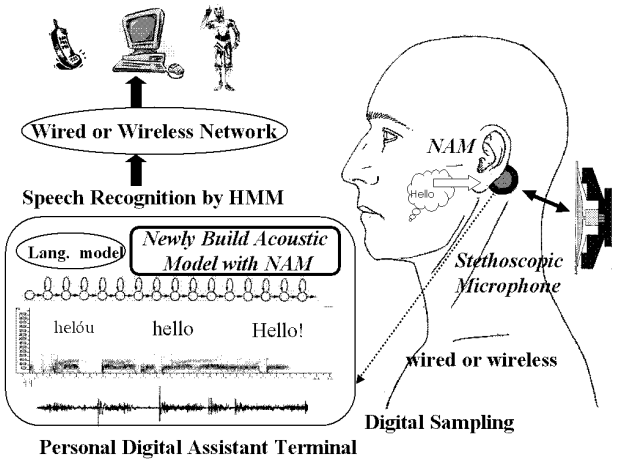

Figure 1. Concept chart of this method

4.3. Development of a new stethoscopic microphone attachment

First we implanted a small condenser microphone into a standard medical-use stethoscope for the purpose of sensing the vibrations of the flesh. Various repeated improvements were performed until we achieved a low-cost microphone, which effectively samples NAM and at the same time is not sensitive to external noise. Currently it is made from the combination of a suction-disc and a polyester plate with one side adhering to the skin. The combination of these two elements forms a micro echoic space and plays an important acoustic role, while helping to adhere the microphone to skin at the same time. Figure 2 shows the cross section and the boom shot of this attachment device. All the component parts can be readily obtained at a do-it-yourself store for simple and low-cost assembly of this lightweight device. We used chemically synthetic rubber that did not bounce for the soundproofing.

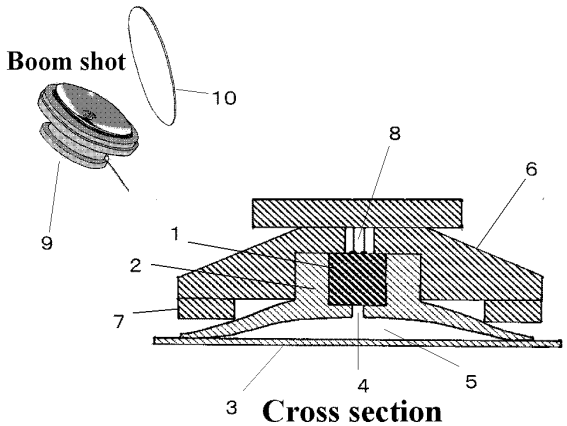

Figure 2. Stethoscopic microphone:

We trained new acoustic models using this NAM sampling from that sampling point on the skin, but the recognition experiments were not successful. Consonant speech sounds were hardly recognized at all, although the vowel sounds were often well recognized when the volume was sufficiently boosted. This was because of the relative difference in power of the consonants (which typically involve strong oral contacts) being too high. On the contrary, when the volume was lowered, the formants of the vowels became too faint to recognize.

Figure 3 shows the speech spectrum and waveform of ordinary speech and whispered speech sampled by an external microphone (top two parts) and NAM speech (bottom) sampled by a stethoscopic microphone attached to the skin on human parotid around the jaw angle. A great difference can be seen in the formant structure.

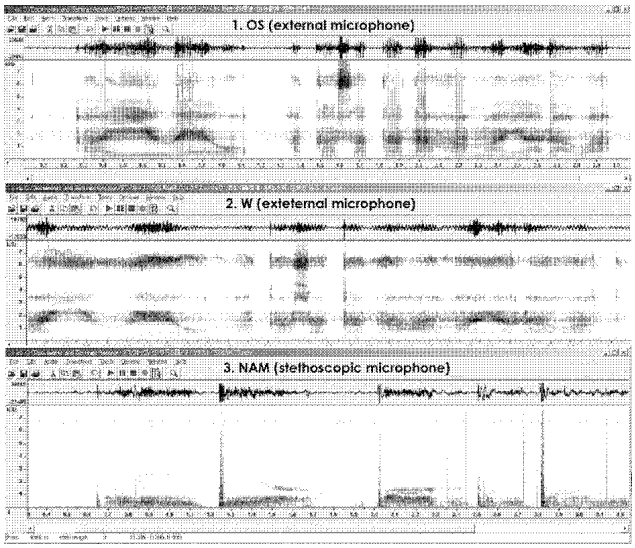

Figure 3. Spectrogram of general NAM compared with ordinary speech (OS) and whisper (W):

4.4. Best sensing position for NAM recognition

We used trial and error to find the best sensing location for the microphone for optimal transmission of non-audible speech sounds through the skin and bone of the head for NAM recognition, balancing the power ratio of vowels and consonants for the best recognition of every phoneme. We finally discovered the point for ideal NAM sensitivity to be just behind the ear. There are hard protruding areas (called mastoid processes) behind the earlobes, which form the terminal of the sternocleido-mastoid muscles. The optimal position for the microphone is with the upper part of its vibration plate just covering the lower part of the mastoid process. From this position we obtain the best levels of NAM for recognition, with an ideal power ratio between vowels and consonants. Figure 4 shows the waveform, spectrum and F0 plot of the ideal NAM. We can't produce F0 contours because of lack of vibration of the vocal folds. There are big differences between the spectra of ideal NAM and those of general NAM obtained from other locations for the microphone attachment.

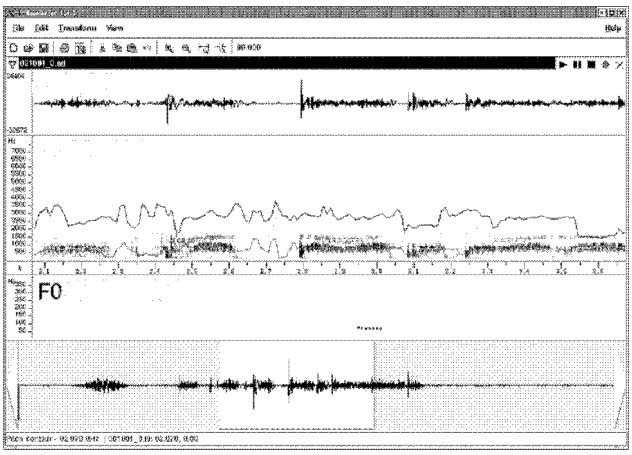

Figure 4. Spectrogram of ideal NAM for recognition:

4.5. Retraining the HMM acoustic models

We trained new HMM acoustic models using ideal NAM recording samples. A male speaker read the ATR 525 phonetically balanced sentences with NAM four times each (the total number of sampling files being 2,100) using HSLab from HTK, with a sampling rate 16KHz [2]. We then used HCopy to convert them to parameterized form into 25^{th}- order (12MFCC+12 Δ MFCC+ Δ power) frames under the same conditions as for ordinary microphone speech. Using these mel-frequency cepstral coefficients (MFCC) as training data with the corresponding label files, we employed HERest to re-estimate male speech models for monophones using 16 Gaussian mixtures. Re-estimation was performed 9 times.

4.6. Results of the Recognition Experiments

For the purpose of evaluation we used the Japanese Dictation Toolkit, a sharable software repository for Japanese Large Vocabulary Continuous Speech Recognition that includes a recognition program together with acoustic and language models. A recognition engine Julius has been developed for the evaluation of both acoustic and language models. These modules can be easily integrated and replaced under a plug-and play framework, which makes it possible to rapidly evaluate components and to develop specific application systems. [4]. Except for changing this ordinary speech acoustic model for the new NAM acoustic model, we used 20K 3-gram language model and decoder with efficient algorithm as provided, without any further changes. With 16KHz sampling recorded NAM raw files we examined NAM recognition results to determine whether we can use it practically or not. As a result we got above 90% accuracy even with the monophone models. And at the same time, we evaluate NAM recognition in acoustically different environments to prove it noise-proof. We confirmed that usual daily life noises such as background music or the sound of TV news did not affect the NAM recognition rates. Evaluations were scored using the scoring scripts provided with the Toolkit. Table 1 shows the output of these scripts. Test No.1 was in a quiet room. No.2 was in a room with background music at the

volume which we usually enjoy (a Bach Concert for Violin and Orchestra). Test No. 3 was with the sound of TV news. We found that NAM recognition can be used comparably with ordinary speech recognition by comparing with the accuracy [4].

Table 1. Results of NAM recognition

SYSTEM SUMMARY PERCENTAGES BY ENVIRONMENT↓

ENV.	Snt	Corr	Acc	Sub	Del	Ins	Err	S.Err
1. QUIET	24	93.61	93.33	4.72	1.67	0.28	6.67	50.00
2. MUSIC	24	91.11	90.00	6.67	2.22	1.11	10.00	62.50
3. TV-NEWS	24	89.72	89.17	9.17	1.11	0.56	10.83	66.67
Sum/Avg	72	91.48	90.83	6.85	1.67	0.65	9.17	59.72

5. PROSODY OF NAM

NAM is a voiceless breath sound, so F0 contour is not plotted as mentioned above. But if the probe of medical ultrasonography is applied vertically on the front surface of the neck, we can observe ups and downs of the whole larynx according to the change of the meter in real time (30 frame/sec). The higher we raise the pitch of the voice, the more up the whole larynx is lifted. Figure5 shows the images of ultrasonography of larynx. The black stripe numbered 3 in the fan-shaped image is the shadow of thyroid cartilage (the Adam's apple). Lower edge of thyroid cartilage is relatively clear as the borderline between black and white. As a calibration, we utter 'do', 'mi', 'sol' and 'do' raised an octave. Line A is the lower edge of thyroid cartilage when 'do' is uttered (first image) and line B is that of 'do' raised an octave (second image). During the speech the lower edge of thyroid cartilage moves between line A and B according to the prosody of it (third image). After capturing animation images to the computer, we cropped one pixel width from a certain height of every frame and merged them by time series. Then we got a contour image as shown the lower part of figure 6 compared with the F0 contour.

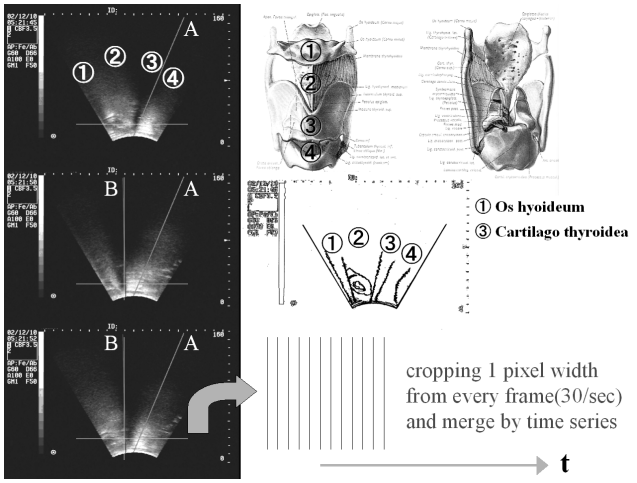

Figure.5 Processing ultrasonographic images of larynx

We propose a new index of the height of larynx standardized by line A and B. We named it Laryngeal Elevation Index (LEI). And the contour showing the up and down movement of larynx (the lower edge of the thyroid cartilage) was named LEI curve as another index of prosody or intention of prosody. Using this LEI curve, we can observe something like prosody of NAM, which is like a speech with a high pitch voice.

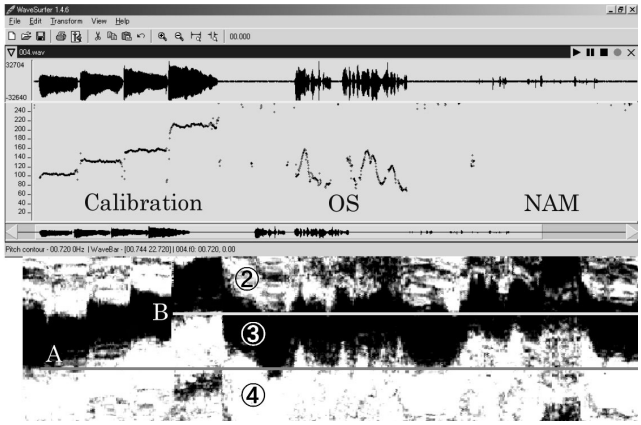

Figure 6. LEI curve, another index of prosody

6. CONCLUSIONS

NAM recognition has been shown to be a practical new speech input interface, being similarly accurate to voiced speech recognition yet robust against noisy environments. The most vital differences from speech recognition for covering wide range of practical use is 'using inaudibly', 'making individual model' and 'robustness against noise'. Use of NAM for speech recognition will result in quieter working environments, not releasing unwanted noises into public spaces such as the office. Furthermore, it will facilitate secure input of speech data into a recognition device, making it more difficult for third-parties to overhear the transactions. We are currently exploring its potential as an input device for vocally handicapped people. This low cost interface of universal design can become a popular hands-free input device for the coming generation of portable or wearable information processing terminals, potentially replacing the keyboard and making third linguistic culture of NAM. Let's free two hands again like an ape had done to be human long ago and build up more civilized world of peace and quiet.

7. REFERENCES

[1] M. Matsuda, H. Mori, and H. Katsuya, "Formant structure of whispered vowels," *J. Acoust. Soc. Jpn.* 56, pp. 447-487, 2000.

[2] S.Yong, J.Jansen, J. Odell D. Ollason, V.Valtchev and Phil Woodland, *The HTK Book*, 2000.

[3] Hideo Suzuki, "Pitfalls when using microphones," *J. Acoust. Soc. Jpn. 55*, pp. 377-381, 1999.

[4] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, M. Yamamoto, A. Yamada, T. Utsuro and K. Shikano, "Overview of Japanese Dictation Toolkit 1999 version" *J. Acoust. Soc. Jpn.* 56, pp. 255-259, 2000.