# Remodeling of the Sensor for Non-Audible Murmur (NAM)

*Yoshitaka Nakajima, Hideki Kashioka*, Kiyohiro Shikano*, Nick Campbell**

ATR Human Information Science Laboratories, Keihanna Science City, 619-0288, Japan
*Graduate School of Information Science, Nara Institute of Science and Technology
yoshi-n@atr.jp

## Abstract

We developed the next generation of skin-attachment sensors for sampling NAM (Non-Audible Murmur) signals by using soft silicone, which has an acoustic impedance close to that of human flesh, as the prime medium of vibration. With new NAM microphones we could sample expanded target voice signal, suppressing air conduction noise signal to low by the experiment of synchronous stereo sampling of air and flesh conduction voices at the same gain. The bandwidth of the soft silicone type NAM microphone has improved and we obtain a much higher accuracy of both NAM and BTOS (Body Transmitted Ordinary Speech) recognition comparėd with the stethoscopic type. Aural comprehension test showed that accuracy of catching sentences of NAM and BTOS has improved with soft silicon type NAM microphone almost as high as that of air conduction voices. However, the extremely low accuracy of meaningless words is a problem to be solved for developing a "Non-Voice Phone".

## 1. Introduction

Non-Audible Murmur (NAM) is a flesh conducted soft whisper which hardly be heard by surrounding people. NAM Microphone is a sensor device for sampling NAM attached to the top of neck skin low behind the earlobe. We have shown the possibility of using NAM signals, i.e., recording quietly articulated breathy speech directly through the flesh, for large vocabulary continuous recognition [1-5] and are working towards an extension of NAM recognition for the wider concept of a "Non-Voice Phone (NAM phone)"; an interface for communication between humans and machines by speech without disturbing surrounding people where the NAM signal is rendered audible by signal processing. For this, a higher sound quality of NAM is needed whether it is digitally processed or not, so we remodeled the NAM microphone.

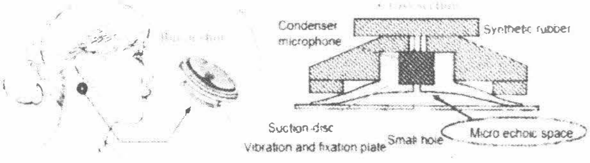

*Figure 1: Stethoscopic Type NAM Microphone*

In our previous work on NAM recognition we used a stethoscopic NAM microphone that we made ourselves. As shown in figure 1, the air in the micro-echoic space between the electrode of an electret condenser microphone (ECM) and a vibration plate play an important role for sensitivity in capturing flesh vibration. Besides NAM, normal speech signal sampled by NAM microphone is defined as Body Transmitted Ordinary Speech (BTOS). But the formants of NAM and BTOS can be observed clearly only below 2 kHz. Amplified NAM sampled with a stethoscopic type microphone sounds like blurred and indistinct whisper and it can be difficult to comprehend certain types of speech especially those containing high band fricative consonants.

## 2. New NAM Microphones using Soft-Silicone

Though the cone shaped micro-echoic space improves the lower band sensitivity of a NAM microphone, we were concerned that this small air-pocket might restrict the bandwidth, so we decided to abolish the air space between the ECM (Electret Condenser Microphone) and the vibration plate. For the sound-carrying medium we chose the type of silicone used for dental molds, which is plastic and easy to shape and form. Since the ECM is designed for sampling air-mediated sound, there is a small air hole on its top surface. We cut the top metal surface of the ECM and exposed the vibration electrode to contact with the vibration media directly as shown in figure 2. We call this an Open-ECM (OECM). We first wrapped the OECM with hard silicone and attached it directly to the skin. Although the sensitivity of this microphone was a little lower than the stethoscopic type, we achieved the wider bandwidth of NAM.

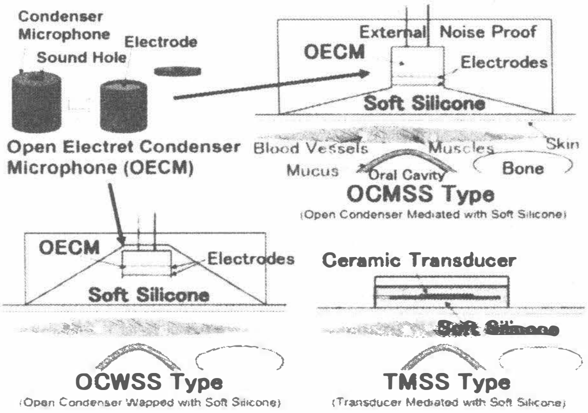

*Figure 2: Soft-Silicone Type NAM Microphone*

However, the more deferent acoustic impedances two substances have, the more sound reflects on their interface. We can observe the structure in the human body using ultrasonography for medical use by taking advantage of this theory of the deference of acoustic impedances. We therefore tested various candidate media by placing them between an ultrasonography probe and the skin. When we tried a piece of soft silicone of the same consistency as human flesh, the inner structure of the body was clearly seen as if there were nothing between probe and skin. For preventing sound loss we found soft silicone better than hard silicone as a medium between the OECM and the skin. Although we would prefer to embed an ECM into the human flesh directly, we can instead wrap an OECM with soft silicone as a good substitute for protruding skin. We therefore tried to replace the air of the cone-shaped micro-echoic space with soft silicone (OCMSS type in figure 2). We made another version, in which a whole OECM was wrapped with soft silicone (OCWSS type in figure 2) because an OECM picks up the sound from every direction and we can separate the OECM from external noise. We can achieve almost the same effect as if the OECM had been implanted into the skin (TMSS type in figure 2). As a substitute of OECM we made another type using disk-shaped ceramic transducer. We made many prototypes of soft silicone NAM microphone, but they can be classified into these three kinds approximately.

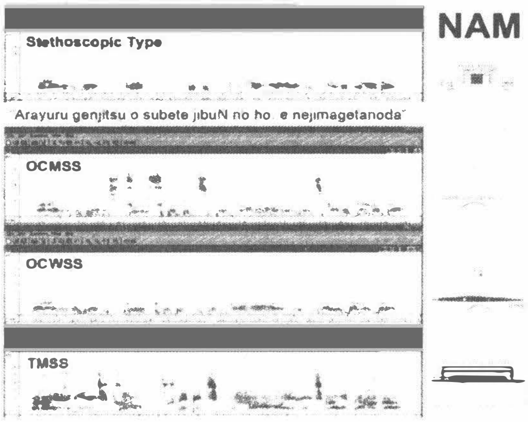

*Figure 3: Spectra of NAM sound*

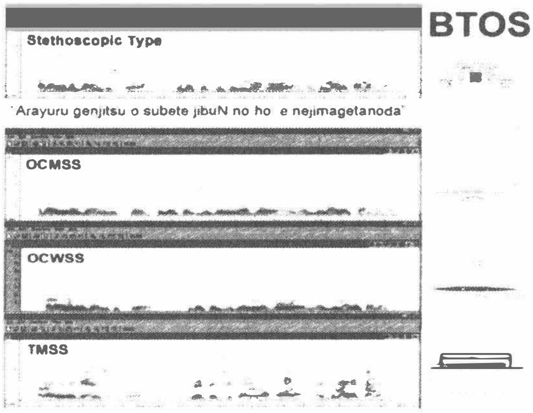

*Figure 4: Spectra of BTOS sound*

Figure 3 shows spectra of NAM sound sampled by 16 kHz with soft-silicone type NAM microphones compared with stethoscopic type. As you see obviously, we obtained much wider bandwidth and clear enough to convey articulated utterances. Figure 4 shows these of BTOS sound spectra sampled with soft-silicone type NAM microphones compared with stethoscopic type.

## 3. Comparison between Air Conduction and Flesh Conduction by Stereo Sampling

The NAM microphone was designed for sampling skin vibrations, so it can be difficult to compare specifications with those of normal microphones for sampling air-conducted sound. In order to show the difference between air conduction and flesh conduction of the voices, we did an experiment of synchronous stereo sampling of both air and flesh conducted voices. As the sensor, ECMs of the same serial number of the same company were used. One of them was processed into the soft silicone type NAM microphone (OCMSS type). The air conduction voice was collected in the distance of 5cm from the mouth like a headset microphone. Right and left tracks were set with the stereo at the same output level and the same amplification rate. As shown in figure 5, in the appropriate gain for NAM sampling the average difference of each frame of power was 5.98 dB. And the average difference was 10.18 dB in the ideal gain for BTOS.

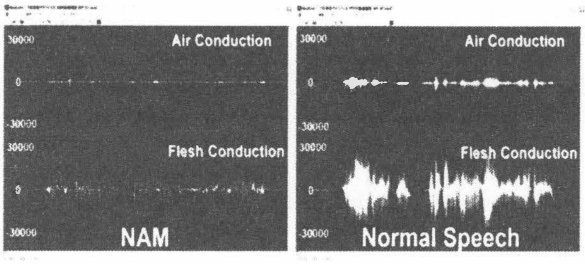

*Figure 5: Comparison between Air Conduction and Flesh Conduction by Stereo Sampling*

In addition we tried the same sampling with background repeated TSP sound as the air conducted external noise. Figure 6 shows the waveforms when input gain was adjusted for NAM and BTOS signal sampling with background repeated TSP sounds. Please pay attention to the amplitude ratio of the target signal and the noise signal. Especially in case of input gain adjusted for BTOS with background repeated TSP sounds, as you can see, the amplitude of TSP sounds got almost invisible in flesh conduction.

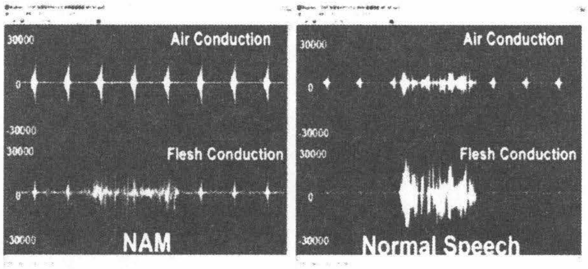

*Figure 6: Comparison between Air Conduction and Flesh Conduction by Stereo Sampling with Iterative TSP Sound*

## 4. Recognition Accuracy by Iterative MLLR

We made NAM and BTOS acoustic model by speaker adaptation (Iterative MLLR), and compared the recognition rates by HMM of the soft silicone type NAM microphone and the stethoscopic type for a numeric evaluation on LVCR.

Figure 7 shows the comparison of the word recognition accuracy by Iterative MLLR [9] of NAM sampled with three types of soft silicone NAM microphone and the stethoscopic type. A male speaker read out 400 newspaper article sentences and 24 sentences for the evaluation by NAM utterance sampled with the same NAM microphone by 16 kHz. The word recognition accuracy was calculated by using JDTK [8] to evaluate the recognition rate. Using HTK [7], we adapted the male normal-speech speaker-independent phonetic tied mixture (PTM) model (64 mixtures and 3000 states) [8] for NAM acoustic model by iterative MLLR of ten times under 128 class regression tree with both mean and variance. Julius3.4 [10] was used as a recognition engine, and 20K dictionary was used for the language model [8]. Word accuracy of the whisper voice recognition with the headset microphone is also ranked in the graph as the contrast on the same condition of iterative MLLR (black bar chart on the right). A horizontal axis is the repetition frequency of adaptation. All three types of soft silicone NAM microphone exceeded the stethoscopic type in word recognition accuracy.

Figure 8 shows the comparison of the word recognition accuracy by the speaker adaptation of BTOS using iterative MLLR on the same condition. The normal speech sampled with the headset microphone is also ranked as the contrast. The contact sensitivity of OCWSS type soft silicone NAM microphone was so high that many BTOS signals saturated to overflow. Therefore, we did not include it into this graph. As for both of soft silicone type NAM microphone, a great improvement was seen about in the BTOS word recognition accuracy compared with the old stethoscope type similarly to NAM recognition.

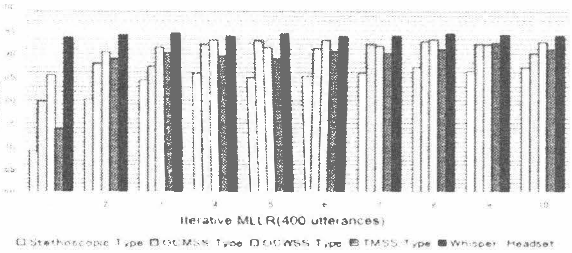

*Figure 7: NAM Recognition Accuracy by Iterative MLLR*

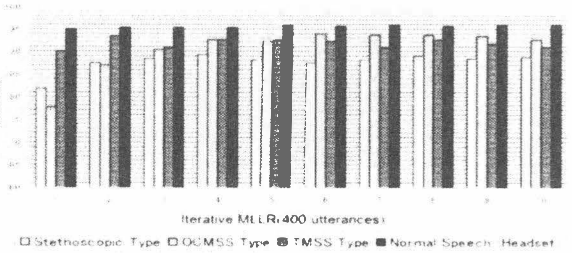

*Figure 8: BTOS Recognition Accuracy by Iterative MLLR*

Word recognition accuracy of the flesh conduction voice is still low compared with the air conduction voice of the contrast as showed in Figure 7 and Figure 8. We think this is caused by whether information on the flesh conduction voice is originally insufficient compared with the air conduction voice or whether the normal-speech speaker-independent model was used as an initial model for speaker adaptation. Or, the both might be the causes. Anyway it is an important theme that the speaker-independent model made from the NAM samples or BTOS samples only. To examine a suitable parameter extraction method for the recognition of the flesh conduction voice is also another attractive theme.

## 5. Hearing Test of Flesh Conduction Voice

We performed hearing test by human using no signal processing NAM and BTOS sound records. Hearing test contained 12 Sentences (daily informational sentence of about ten words or more), 12 single words and 12 meaningless words (jargon of 3-4 moras). 12 volunteers of each generation from teens to sixties (6 male and 6 female) took the test. We prepared 12 kinds of recording according to the kind of speech (NAM or BTOS), sensing device (soft silicone type or stethoscopic type) and sampling rate (8 kHz or 16 kHz), containing normal speech and whisper sampled with headset microphone as control. Sentences and single words were selected carefully so that even teens understood enough. We made random permutation of 12 kinds of recordings at each line of question matrix, so 12 volunteers heard 12 kinds of recording randomly. Every volunteer was supposed to listen to each utterance three times.

Independently of kinds of speech and sensing devices, there were no significant differences in hearing recognition accuracy between 8 kHz sampling and 16 kHz sampling.

Figure 9 shows the comprehension accuracy of sentence utterances by NAM. At all hearing, soft silicon type NAM microphone showed significantly high recognition accuracy compared with stethoscopic type. Differences of recognition accuracy between NAM sampled with soft silicone type NAM microphone and air conducted whisper sampled with headset microphone were not significant at second and third hearing.

And figure 10 shows the accuracy of sentence utterances by BTOS. At all hearing, soft silicon type NAM microphone showed significantly high recognition accuracy compared with stethoscope type. Differences of recognition accuracy between BTOS sampled with soft silicone type NAM microphone and air conducted normal speech sampled with headset microphone were not significant at all hearing.

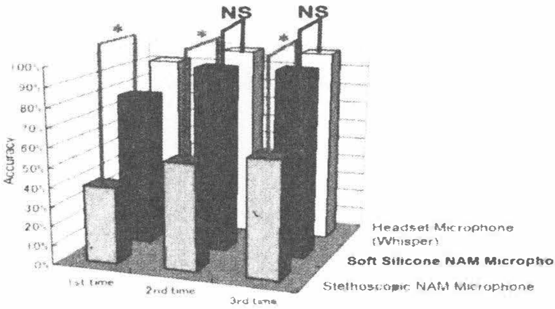

*Figure 9: Hearing Accuracy of Sentence (NAM)*

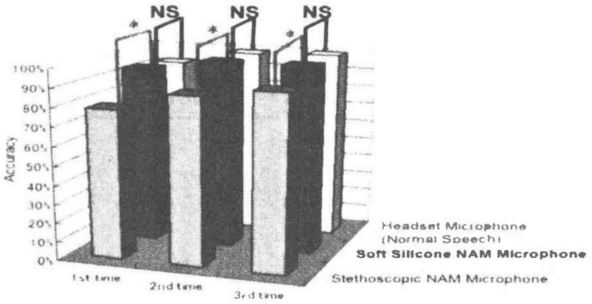

*Figure 10: Hearing Accuracy of Sentence (BTOS)*

On the left of figure 11 shows the accuracy of single word utterances by NAM. Similar tendency was seen to the case of sentence utterances. But as for meaningless words on the right, even air conducted whisper showed below 50% recognition rate and NAM utterance was hardly caught and recognition accuracy was extremely low.

And on the left of figure 12 shows the accuracy of single word utterances by BTOS. Similar tendency was seen to the case of sentence utterances. Even in the case of meaningless words normal speech sampled with headset microphone was highly caught almost 100% unlike air conducted whisper utterances. And recognition accuracy of BTOS was below 50%, though it was higher than that of NAM.

Although soft silicone NAM microphone improved the hearing accuracy in BTOS and NAM in comparison with stethoscopic type, the catching accuracy of flesh conduction voice was considerably inferior to the air conduction voice in the situation by which the phoneme should be caught like a meaningless word especially in the case of NAM utterance.

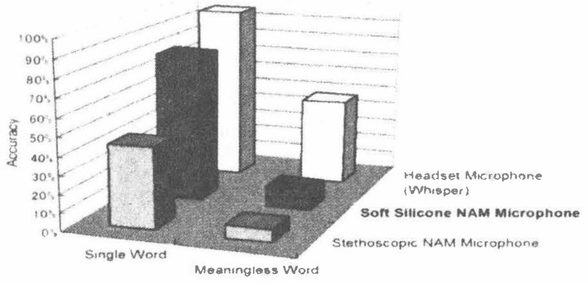

*Figure 11: Hearing Accuracy of Single Word (NAM)*

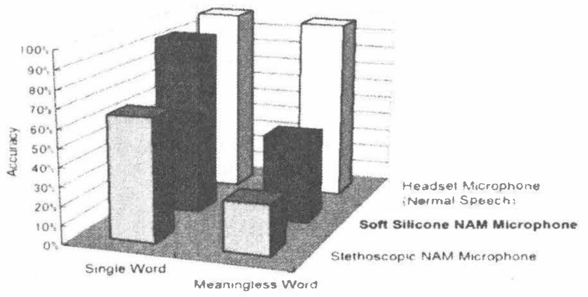

*Figure 12: Hearing Accuracy of Single Word (BTOS)*

## 6. Conclusion

We developed the next generation of skin-attachment sensors for sampling NAM signals by using soft silicone, which has the acoustic impedance close to that of human flesh. These new NAM microphones enable us to record more wideband NAM sound, which is clear enough to convey articulated utterances even without digital signal processing. We obtain a much higher accuracy of recognition by HMM in both NAM and BTOS compared with former stethoscopic type. And new NAM microphone improves the human hearing accuracy in both NAM and BTOS. We believe that not only NAM recognition but also "Non-Voice Phone (NAM phone)" will be possible. But the hearing accuracy of NAM is considerably inferior to the normal voice when each phoneme should be caught like meaningless jargons. That's an inevitable problem how to handle unknown words and it will play an important role to modulate NAM to audible voice by signal processing. We propose NAM signal as a new all-purpose voice input interface and present speech signal-processing algorithms that allow this speech source to be used for human-to-human and human-to-machine communication interface which is robust to noisy environments yet unobtrusive even in a quiet room where other people may be present.

## 7. References

[1] Y. Nakajima, H. Kashioka, K. Shikano, N. Campbell: "Non-Audible Murmur Recognition Input Interface Using Stethoscopic Microphone Attached to the Skin," Proc. ICASSP, pp.708-711, 2003.

[2] Y. Nakajima, H. Kashioka, K. Shikano, N. Campbell: "Non-Audible Murmur Recognition," Proc. EUROSPEECH, pp.2601-2604, 2003.

[3] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, K. Shikano: "Accurate Hidden Markov Models for Non- Audible Murur (NAM) Recognition Based on Iterative Supervised Adaptation," Proc. ASRU, pp.171-185, 2003.

[4] P. Heracleous, Y. Nakajima, A. Lee, Hiroshi H. Saruwatari, K. Shikano: "Audible (Normal) Speech and Inaudible Murmur Recognition Using NAM Microphone," Proceedings of the 12th European Signal Processing Conference (EUSIPCO2004), pp.329-332, Sept. 2004.

[5] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, K. Shikano: "Non-Audible Murmur (NAM) Speech Recognition Using a Stethoscopic NAM Microphone," Proceedings of 8th International Conference on Spoken Language Processing (ICSLP2004), WeC2102p-6, pp.527-530, October 2004.

[6] S. C. Jou, T. Schultz, A. Waibel: "Adaptation for Soft Whisper Recognition Using a Throat Microphone," Proc. 8th International Conferences on Spoken Language Processing (ICSLP2004), WeC2102p.12, Oct. 2004.

[7] S. Yong, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland: The HTK Book (for HTK Version 3.2.1), Cambridge University Engineering Department, 2002.

[8] 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.

[9] P. C. Woodland, D. Pye and M.J.F. Gales: "Iterative Unsupervised Adaptation Using Maximum Likelihood Linear Regression," Proc. ICSLP, pp.1133- 1136, 1996.

[10] A. Lee, T. Kawahara, K. Shikano: "Julius - An Open Source Real-Time Large Vocabulary Recognition Engine," Proc. 7th European Conference on Speech Communication and Technology (EUROSPEECH2001), pp.1691-1694, 2001.