Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 94068, 11 pages
doi:10.1155/2007/94068

Research Article

Unvoiced Speech Recognition Using Tissue-Conductive Acoustic Sensor

Panikos Heracleous,[1,2] Tomomi Kaino,[1] Hiroshi Saruwatari,[1] and Kiyohiro Shikano[1]

[1] *Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma-shi, Nara 630-0192, Japan*
[2] *Department of Computer Science, University of Cyprus, 75 Kallipoleos Street, P.O. Box 537, 1678 Nicosia, Cyprus*

Received 22 September 2005; Revised 6 January 2006; Accepted 30 January 2006

Recommended by Matti Karjalainen

We present the use of stethoscope and silicon NAM (nonaudible murmur) microphones in automatic speech recognition. NAM microphones are special acoustic sensors, which are attached behind the talker's ear and can capture not only normal (audible) speech, but also very quietly uttered speech (nonaudible murmur). As a result, NAM microphones can be applied in automatic speech recognition systems when privacy is desired in human-machine communication. Moreover, NAM microphones show robustness against noise and they might be used in special systems (speech recognition, speech transform, etc.) for sound-impaired people. Using adaptation techniques and a small amount of training data, we achieved for a 20 k dictation task a 93.9% word accuracy for nonaudible murmur recognition in a clean environment. In this paper, we also investigate nonaudible murmur recognition in noisy environments and the effect of the Lombard reflex on nonaudible murmur recognition. We also propose three methods to integrate audible speech and nonaudible murmur recognition using a stethoscope NAM microphone with very promising results.

1. INTRODUCTION

The NAM microphone [1] belongs to the acoustic sensor paradigm, in which speech is conducted not through the air, but within body tissues, bone, or the ear canal. The NAM microphone is attached behind the talker's ear and speech is captured through body tissue. Figure 1 shows the attachment of a NAM microphone to the talker.

The bone-conductive microphone used in [2, 3], the throat microphone used in [4], and the ear-plug used in [5] are acoustic sensors similar to NAM microphones. Basically, in those studies a nonconventional acoustic sensor combined with a standard microphone was used to increase the robustness against noise. In [6] a prototype stethoscope NAM microphone and a throat microphone were used for soft whisper recognition in a clean environment.

NAM microphones are special acoustic sensors, which can capture not only normal (audible) speech, but also very quietly uttered speech (nonaudible murmur). As a result, NAM microphones can be applied in automatic speech recognition systems when privacy is desired in human-machine communication. Moreover, since a NAM microphone receives the speech signal directly from the body, it shows robustness against the environmental noises. In addition, it might be also used in special systems (speech recognition, speech transform, etc.) for sound-impaired people.

The stethoscope microphone is based on stethoscopes used by medical doctors to examine the patients. In a very similar device, a microphone is used covered by a membrane. On the other hand, the silicon microphone uses a microphone wrapped by silicon. The idea to use silicon is based on the fact that silicon has similar impedance to that of human flesh.

Our current research, focuses on the recognition of nonaudible murmur using NAM microphones in various environments. Previously, in [7] speaker-dependent nonaudible murmur recognition in a clean environment and using a stethoscope NAM microphone was reported. In that work, context-independent hidden Markov models (monophones) and expectation-maximization (EM) training procedure were used. To evaluate the performance of nonaudible murmur recognition using context-dependent models, we conducted experiments using phonetic tied mixture (PTM)

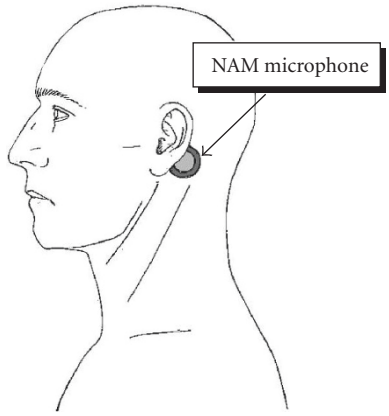

Figure 1: NAM microphone attached to the talker.

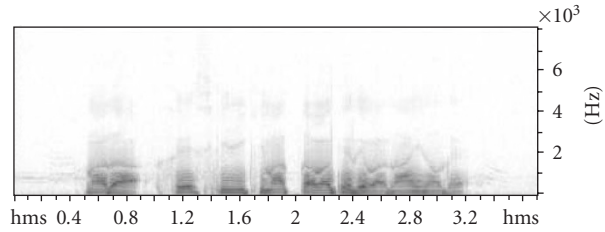

Figure 2: Spectrogram of an audible Japanese utterance captured by a NAM microphone.

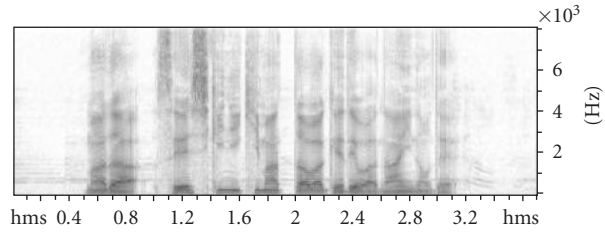

Figure 3: Spectrogram of an audible Japanese utterance captured by a close-talking microphone.

models [8] and stethoscope and silicon NAM microphones. However, instead of EM training procedure we applied speaker-adaptation techniques, which require significantly less amount of training data [9–11]. The achieved results are very promising and show the effectiveness of applying adaptation methods for nonaudible murmur recognition. Using a small amount of training data, we achieved for a 20 k dictation task a 93.9% word accuracy for nonaudible murmur recognition in a clean environment. Following the previous works, in [12] they also conducted experiments using a silicon NAM microphone and applying adaptation techniques with similar results. In addition to the experiments in a clean environment, we also carried out experiments using clean models and noisy test data [13].

To make a nonaudible murmur-based speech recognition system more flexible, we conducted experiments for audible speech and nonaudible murmur recognition using a stethoscope NAM microphone. The achieved results show the effectiveness of a NAM microphone in an integrated normal-speech and nonaudible-murmur recognition system [14].

In this paper, we also investigate the NAM microphone robustness against noise using simulated and real noisy data. We also conducted experiments using Lombard nonaudible murmur data, showing that the Lombard reflex affects nonaudible murmur recognition markedly.

2. NONAUDIBLE MURMUR CHARACTERISTICS

Nonaudible murmur and audible speech captured by a NAM microphone have different characteristics compared with air-conducted speech. Similarly to whisper speech, nonaudible murmur is unvoiced speech produced by vocal cords not vibrating and does not incorporate any fundamental (F0) frequency. Moreover, body tissue and loss of lip radiation act as a low-pass filter and the high-frequency components are attenuated. However, the nonaudible murmur spectral components still provide sufficient information to distinguish and recognize sounds accurately.

Figure 2 shows the spectrogram of an audible Japanese utterance captured by a stethoscope NAM microphone and Figure 3 shows the spectrogram of the same utterance captured by a close-talking microphone. Both figures show that the utterance captured by a NAM microphone is of limited frequency band, namely, it contains frequency components up to 3–4 kHz.

Due to these differences, normal-speech hidden Markov models (HMMs) cannot be used for recognition of speech captured by a NAM microphone. To realize nonaudible murmur recognition, new HMMs have to be trained using nonaudible murmur database.

3. NONAUDIBLE MURMUR AUTOMATIC RECOGNITION

In this section, we present experimental results for speaker-dependent nonaudible murmur recognition using NAM microphones. The recognition engine used was the Julius 20 k vocabulary Japanese dictation toolkit [15]. The recognition task was large vocabulary continuous speech recognition. A trigram language model trained with newspaper articles was used. The perplexity of the test set was 87.1. The initial models were speaker-independent, gender-independent, 3000-state phonetic PTM HMMs, trained with the JNAS database [16] and the feature vectors were of length 25 (12 MFCC (mel-frequency cepstral coefficients), 12 Δ MFCC, Δ E). Table 1 shows the system specifications.

The nonaudible murmur HMMs were trained using a combination of supervised 128-class regression tree MLLR [17] and MAP [18] adaptation methods. Using, however, the MLLR and MAP combination, the parameters are initially transformed using MLLR, and the transformed parameters

Table 1: System specifications.

Sampling frequency	16 kHz
Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order Δ MFCCs 1-order Δ E
HMM	PTM, 3000 states
Training data	JNAS/nonaudible murmur
Test data	nonaudible murmur

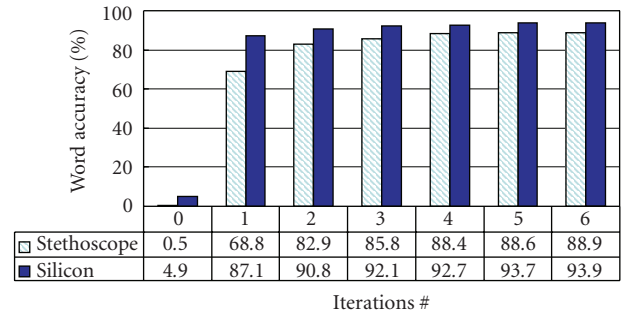

Figure 4: Nonaudible murmur recognition in a clean environment.

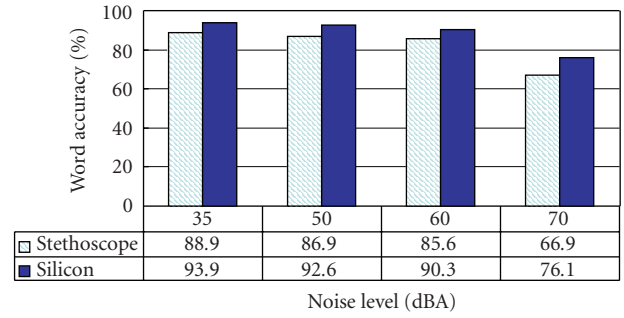

Figure 5: Nonaudible murmur recognition in noisy environments (superimposed noisy data).

are used as priors in MAP adaptation. In this way, during MLLR the acoustic space is shifted and the MAP adaptation performs more accurate transformations. Moreover, due to the use of a regression tree in MLLR, parameters which do not appear in the training data, and therefore are not transformed during MAP, are transformed initially during MLLR.

Due to the large difference between the training data and the initial models, single-iteration adaptation is not effective in nonaudible murmur recognition. Instead, a multi-iteration adaptation scheme was used. The initial models are adapted using the training data and the intermediate adapted models were trained. The intermediate models were used as initial models and were re-adapted using the same training data. This procedure was continued until no further improvement was obtained. Results show, that after 5–6 iterations significant improvement was achieved compared with the single-iteration adaptation. This training procedure is similar to that proposed by Woodland et al. [19], but the object is different.

3.1. Experiments using clean and simulated noisy test data

In this experiment, both training and test data were recorded in a clean environment by a male speaker using NAM microphones. For training 350 and for testing 48 nonaudible murmur utterances of a male speaker were used. Figure 4 shows the achieved results. As the figure shows, the results are very promising. Using a small amount of data and adaptation techniques, we achieved high word accuracies. More specifically, using a stethoscope microphone we achieved an 88.9% word accuracy and using a silicon NAM microphone we achieved a 93.9% word accuracy for nonaudible murmur recognition. The results also show the effect of the multi-iteration adaptation scheme. As can be seen, with increasing number of adaptation iterations, the word accuracy was markedly increased.

We also conducted an experiment using simulated noisy data. In this experiment, the same clean 350 utterances were used for adaptation. For testing, 48 noisy nonaudible murmur utterances were used. Noise recorded in an office was played back at 50 dBA (decibels adjusted), 60 dBA, and 70 dBA levels and was recorded using NAM microphones. The recorded noises were superimposed onto the clean data to create the noisy test data.

Figure 5 shows the obtained results. As can be seen, for the 50 dBA and 60 dBA noise levels the performance was almost equal to that of the clean case. When the noise level became 70 dBA, the performance decreased, however, still nonaudible murmur recognition with reasonable results was possible. Note, that no additional noise reduction approaches were used, and that the HMMs were trained using clean data. Results show that stethoscope NAM microphone is less robust against noise, particularly at the 70 dBA noise level.

Figure 6 shows the long-term spectrum of the noise used in our experiments. Noises captured by NAM microphones were superimposed onto the clean test data to simulate the noisy test data. Figure 7 shows the spectrum of the noise recorded using NAM microphones at 70 dBA level. The figure shows the similarity in the spectra of the two captured noises. Differences appear between 3 kHz and 5 kHz, where noise captured by the stethoscope microphone shows a higher spectral content. This might explain the significant decrease in word accuracy at 70 dBA when using the stethoscope microphone.

3.2. Experiments using real noisy test data

In this subsection, we report experimental results for nonaudible murmur recognition using real noisy database.

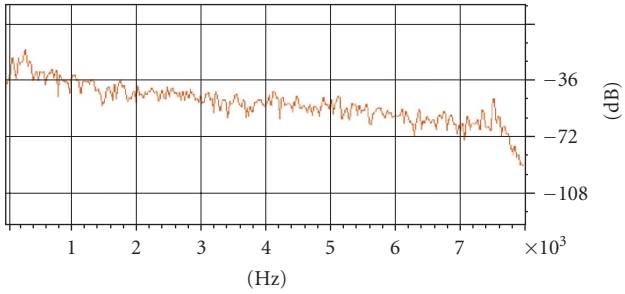

Figure 6: Long-term power spectrum of office noise used in the experiments.

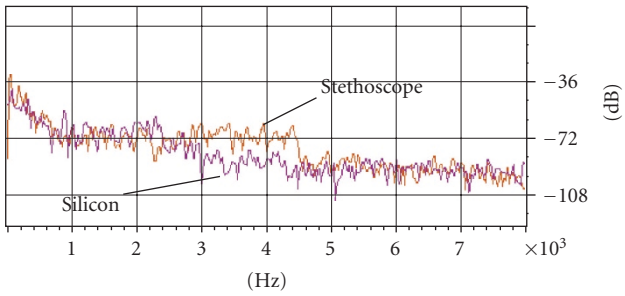

Figure 7: Long-term power spectrum of office noise at 70 dBA level captured by NAM microphones.

The noisy test data were recorded in an environment using a silicon NAM microphone, where different types of noise were playing back at 50 dBA and 60 dBA levels, while a female speaker was uttering the test data. Four types of noise were used (office, car, poster-presentation, and crowd). For each noise and each level 24 utterances were recorded. For adaptation 100 clean utterances were used. For comparison, we also created superimposed noisy data using the same clean test utterances and office noise captured using a silicon NAM microphone. The speaker in this experiment was different than the speaker in the previous experiments.

Figure 8 shows the obtained results when using office noise in comparison with the case when the same noise was superimposed on the clean data. As can be seen, using real noisy test data, the performance decreases. Namely, at the 50 dBA noise level the obtained word accuracy was 68.4% and at the 60 dBA noise level 46.5%.

Figure 9 shows the word accuracies for the four types of noise. The results are similar to the previous ones. With increasing noise level, word accuracy decreases significantly. For the clean case we achieved an 83.7% word accuracy, for the 50 dBA noise level a 66.9% word accuracy on average, and for the 60 dBA noise level a 53.3% word accuracy on average. In the case of car and crowd noises, the difference between the 50 dBA and 60 dBA performances is not very large. In the case of poster-presentation and office noises, the difference is larger.

Although the performance using real noisy data is not markedly low and nonaudible recognition is still possible,

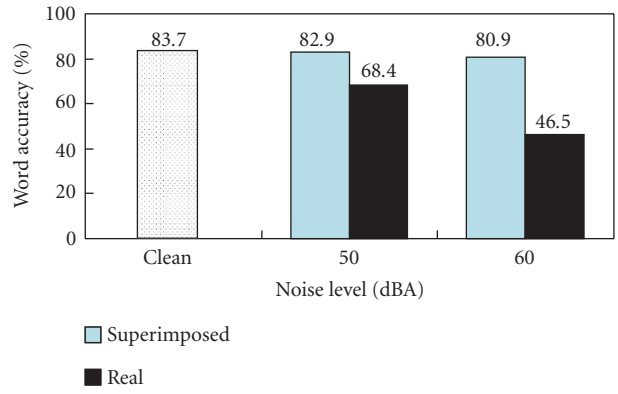

Figure 8: Nonaudible murmur recognition using noisy test data (office noise).

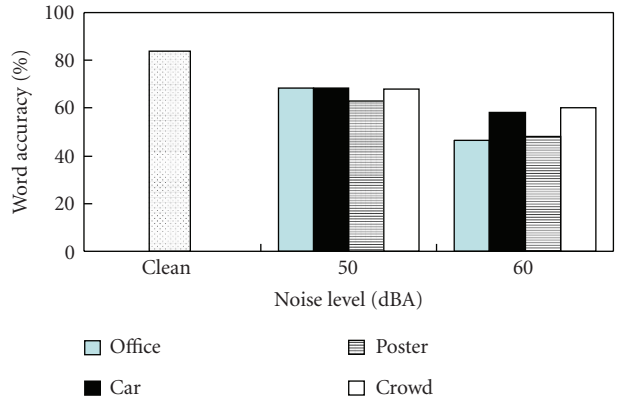

Figure 9: Nonaudible murmur recognition using various types of noise.

further investigations are necessary. In several studies, a negative impact effect of the Lombard reflex [20–24] on automatic recognizers for normal speech has been reported. It is possible, therefore, that the degradations in word accuracy for nonaudible murmur recognition when using real noisy data are also related to the Lombard reflex. To realize this, we also addressed the Lombard reflex problem.

4. THE ROLE OF THE LOMBARD REFLEX IN NONAUDIBLE MURMUR RECOGNITION

When speech is produced in noisy environments, speech production is modified leading to the Lombard reflex. Due to the reduced auditory feedback, the talker attempts to increase the intelligibility of his speech, and during this process several speech characteristics change. More specifically, speech intensity increases, fundamental frequency (F0) and formants shift, vowel durations increase and the spectral tilt changes. As a result of these modifications, the performance of a speech recognizer decreases due to the mismatch between the training and testing conditions.

To show the effect of the Lombard reflex, Lombard speech is usually used, which is a clean speech uttered

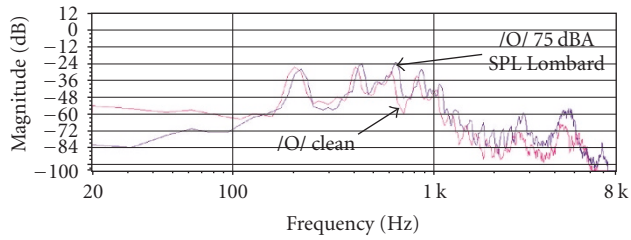

Figure 10: Power spectrum of clean vowel /O/ and Lombard vowel /O/.

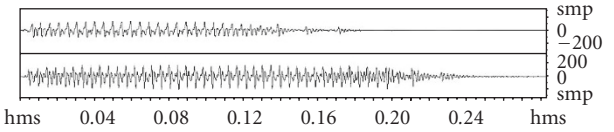

Figure 11: Waveform of clean vowel /O/ (upper) and Lombard vowel /O/.

while the speaker listens to noise through headphones or earphones. Though Lombard speech does not contain noise components, modifications in speech characteristics can be realized.

Figure 10 shows the power spectrum of a normal-speech clean vowel /O/ and a Lombard vowel /O/ recorded while listening to office noise through headphones at 75 dBA noise level. The figure clearly shows the modifications leading to the Lombard reflex; power increased, formants shifted, and spectral tilt changed. Figure 11 shows the waveforms of the clean and Lombard /O/ vowels. As can be seen, the duration and amplitude of the Lombard vowel also increased. These differences in the spectra cause feature distortions (e.g., mel-frequency cepstral coefficients (MFCC) distortions), and acoustic models trained using clean speech might fail to correctly match speech affected by the Lombard reflex.

Figure 12 shows the waveform, spectrogram, and FO contour of a Lombard nonaudible utterance recorded at 80 dBA using a silicon NAM microphone. The figure shows the effect of the Lombard reflex. Although the speaker attempts to speak in nonaudible murmur manner, due to the presence of noise his speech becomes voicing with vocal cords vibrating. As can be seen, this Lombard speech has characteristics similar to those of normal speech (e.g., pitch, formants, etc.) and differs from nonaudible murmur. Therefore, when nonaudible murmur recognition is performed in noisy environments, the produced nonaudible murmur characteristics are different than those of the nonaudible murmur used in the training. As a result, the performance is degraded, even though the NAM microphone can capture nonaudible murmur without a high sensitivity to environmental noise.

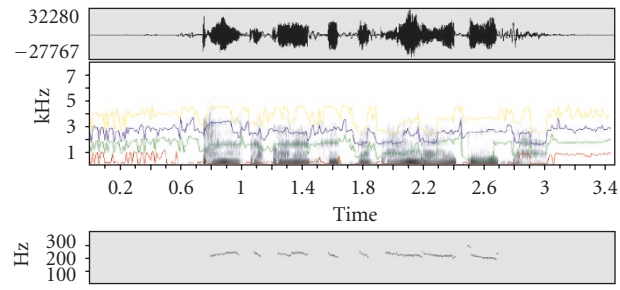

Figure 12: Lombard nonaudible murmur recorded at 80 dBA.

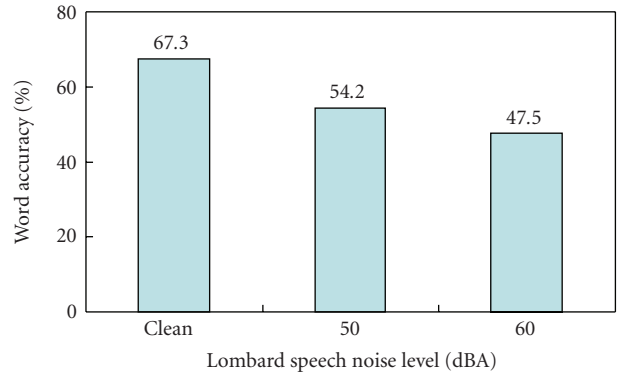

Figure 13: Nonaudible murmur recognition using Lombard data.

4.1. Experiment showing the effect of the Lombard reflex on nonaudible murmur recognition

To show the effect of the Lombard reflex on nonaudible murmur recognition, we carried out a baseline experiment using Lombard nonaudible murmur test data recorded using a silicon NAM microphone. The data were recorded in an anechoic room, while the speaker was listening to office noise through headphones. Since we used high-quality headphones, we assumed that no noise from the headphones was added to the recorded data. We recorded 24 clean utterances, 24 utterances at 50 dBA, and 24 utterances at 60 dBA noise levels. The acoustic models used were trained with clean nonaudible murmur data using 50 utterances and MLLR adaptation. The data were uttered by a female speaker, other than the previous ones.

Figure 13 shows the obtained results and the effect of the Lombard reflex on nonaudible murmur recognition. Using clean test data, we achieved a 67.3% word accuracy, using 50 dBA Lombard data a 54.2% word accuracy, and using 60 dBA Lombard data a 47.5% word accuracy. These results show an analogy between the experiments using real noisy data and the experiment using Lombard data. In both cases, the performances decreased almost equally.

In nonaudible murmur phenomena, the Lombard reflex is also present when there is no masking noise. However, due to the very low intensity of nonaudible murmur, speakers might not hear their own voice. To make their voice audible,

Table 2: Lombard nonaudible murmur recognition using matched and crossed models.

HMM level [dBA]	Training level [dBA]		
	50	60	70
50	64.8	51.1	45.4
60	64.2	65.8	39.5
70	50.0	39.8	72.9

they increase their vocal levels, and as a result, nonaudible murmur changes to voicing.

4.2. Lombard nonaudible murmur recognition using matched and crossed HMMs

In this experiment, we further investigate the recognition of Lombard nonaudible murmur. Our final aim is to increase the word accuracies of nonaudible murmur recognition in real noisy environments taking also into account the Lombard reflex and incorporating Lombard reflex characteristics in creating acoustic models for nonaudible murmur. Therefore, as a first step we conducted experiments using matched and crossed HMMs, and we propose the training of a multilevel Lombard nonaudible murmur HMMs set for recognition of arbitrary Lombard-level test data.

We trained acoustic models using MLLR and nonaudible murmur data of 50 dBA, 60 dBA, and 70 dBA Lombard level (e.g., level of the noise which hears the talker through headphones while uttering the data. The data do not contain any noise). For training, we used 50 utterances for each level and for testing 24 utterances for each level. Table 2 shows the achieved results. The results show that using matched models the word accuracy increases with increasing the Lombard noise level. With increasing the noise level, however, the talker attempts to increase the intelligibility of his speech and as a result the quality of Lombard nonaudible murmur becomes higher. On the other hand, the results show the difficulties in recognizing Lombard nonaudible murmur using acoustic models trained with other Lombard-level data. With increasing the Lombard level, the mismatch between nonaudible murmurs also increases and word accuracies decrease.

For recognition of Lombard nonaudible murmur of various levels, we applied a method based on multi-level Lombard HMMs. More specifically, we trained a common HMMs set using the whole training data (clean, 50 dBA, 60 dBA, and 70 dBA) and we recognized the various Lombard-level test utterances. Figure 14 shows the achieved results. Using only a common HMMs set, we recognized arbitrary Lombard utterances with a 74.9% word accuracy on average. Moreover, in the cases of 50 dBA and 60 dBA Lombard levels the word accuracies are even higher compared with those of the matched cases. However, due to the low mismatch between 50 dBA and 60 dBA Lombard levels the training of a common HMMs set with more data has the same effect as if we increase the adaptation data in the matched cases.

In this experiment, using 24 clean test utterances the recognition accuracy was 82.7% when using clean models (e.g., matched models trained with 50 clean utterances).

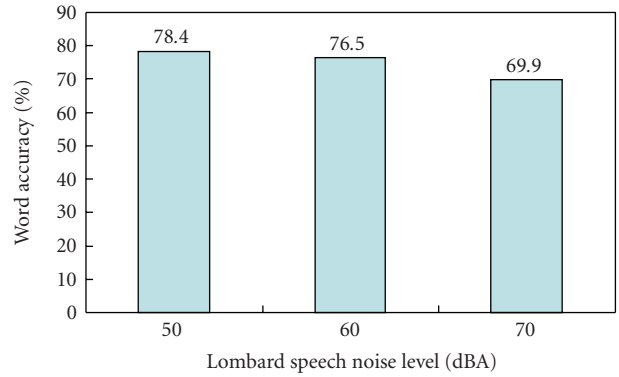

Figure 14: Nonaudible murmur recognition using Lombard data and a multilevel common HMM set.

5. AUDIBLE SPEECH RECOGNITION USING A STETHOSCOPE MICROPHONE

The achieved results show the effectiveness of NAM microphone in nonaudible murmur recognition, though we conducted only speaker-dependent experiments and we used a relatively small test set. Using a NAM microphone and a small amount of adaptation data, we recognized speech uttered very quietly with very high accuracy. NAM microphones can be used as a part of a recognition system, when privacy in communication is very important (e.g., telephone speech recognition applications). A NAM-based speech recognition system, however, has limited applications. Moreover, it requires a special and, less user friendly way in human-machine communication, which is not always necessary. For practical reasons, the system should be also able to recognize audible speech.

In this section, we also focus on this problem, and we show that NAM microphone can be used for audible speech recognition, taking also advantage of its robustness against noise.

Figure 15 shows the waveform of a normal-speech signal received by a close-talking microphone. Figure 16 shows the same signal received by a NAM microphone. The two signals are synchronized, due to a two-channel recording. The figures show the high similarity between the two signals. Figures 17 and 18 show the spectra of the received speech signals. As can be seen, the spectra show similarities up to 1 kHz. After 1 kHz the NAM spectral components are attenuated and from 3 kHz remain flat. As a result of the high-frequency attenuation, the quality of the signal received by the NAM microphone is lower. Figures 19 and 20 show the F0 contours of the previously described signals, which are very similar.

The different frequency characteristics of the two signals require different approach for speech recognition. More specifically, the acoustic models used to recognize audible speech received by a close-talking microphone cannot be used for recognition of normal speech received by a NAM microphone. Therefore, it is necessary to train a new acoustic models set.

The HMM set for recognition of audible speech received by NAM microphone was created using iterative MLLR. A

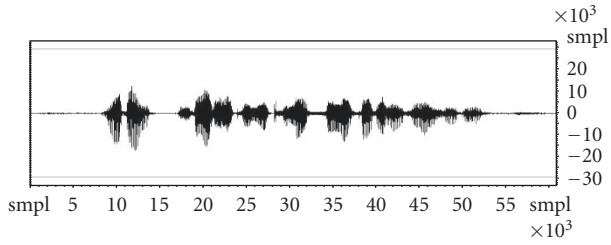

Figure 15: Normal speech waveform—close-talking microphone.

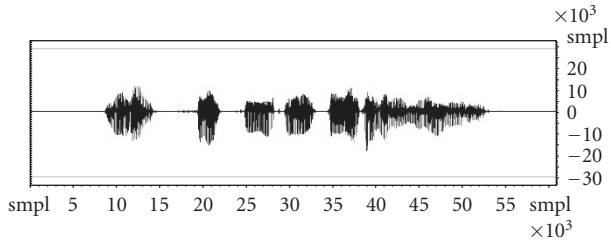

Figure 16: Normal speech waveform—NAM microphone.

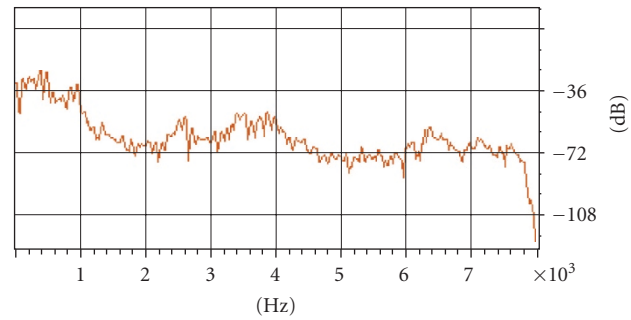

Figure 17: Normal speech long-term spectrum—close-talking microphone.

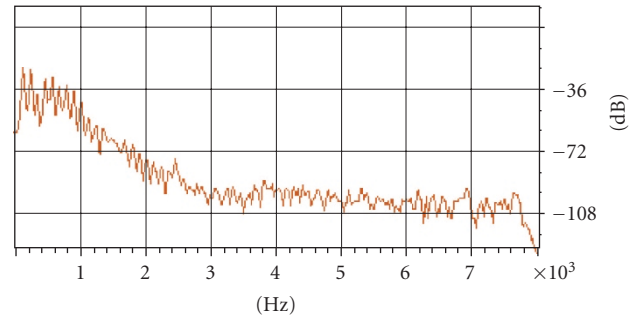

Figure 18: Normal speech long-term spectrum—NAM microphone.

128-class regression tree, 350 adaptation utterances, and 4 iterations were used. For evaluation, 72 NAM utterances recorded under several conditions (quiet, background music, TV-news) were used. For comparison, we trained HMMs for recognition of normal speech received by a close-talking microphone. Single-iteration MLLR with 32-class regression tree, and 100 adaptation utterances were used. The adaptation parameters and the adaptation amount were adjusted after conducting several experiments to select the optimal ones.

Table 3 shows the achieved results. As can be seen, in quiet environment the speech received by NAM microphone was recognized with slightly lower accuracy. The reason is that the spectral content is lost during tissue transmission. In the case, however, when there is a background noise (music, TV-news) the recognition of audible speech received by NAM microphone showed higher performance. Although under noisy environments the performance decreased, we observe that the decreases are not significant. More specifically, in a quiet environment we achieved 93.8% word accuracy, and in noisy environments 93.2% and 92.9%, respectively. The achieved results show the effectiveness of a NAM microphone for audible speech recognition. Especially, in noisy environments this is a very important advantage.

6. INTEGRATED AUDIBLE (NORMAL) SPEECH AND NONAUDIBLE MURMUR

A challenging topic is to integrate audible and nonaudible murmur recognition. In the previous sections, we showed the effectiveness of a NAM microphone in nonaudible murmur and audible speech recognition. A recognition system, which combines recognition of the two types of speech using a NAM microphone, can be very flexible and practical. However, in cases when privacy is not important, user can talk in a normal manner. On the other hand, users can communicate with a speech recognition-based system in a way that other listeners cannot hear their conversation. In this section, we introduce three techniques to integrate nonaudible murmur and audible speech recognition. These approaches are based on case-dependent HMMs created using iterative MLLR and training data recorded using a stethoscope NAM microphone.

6.1. *Gaussian mixture models (GMMs) based discrimination*

The first approach is based on GMM-based discrimination. Two GMMs (one-emitting state HMM) were trained using audible speech and nonaudible murmur received by a NAM microphone, respectively. The transcriptions of the uttered speech were merged to form only one model. Figure 21 shows the block diagram of the system. A NAM microphone is used to receive the uttered speech. After analysis, matching is performed between the input speech and the two GMMs. The matching provides a score for each GMM. These scores are used by the system to make decision about the input speech. Then, the system switches to the corresponding HMMs and speech recognition is performed in a conventional way. The HMMs sets used in this experiment are the same as in the experiments described in Section 4. To evaluate the performance of the method, we carried out

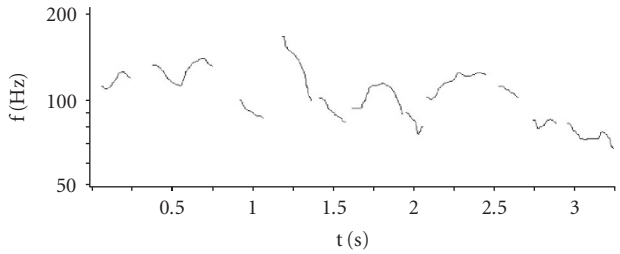

Figure 19: F0 contour of normal speech—close-talking microphone.

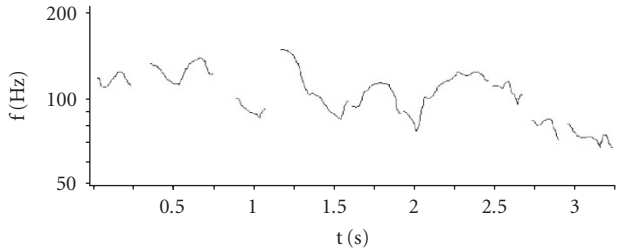

Figure 20: F0 contour of normal speech—NAM microphone.

Table 3: Recognition rates for audible speech.

Microphone	Word Accuracy (%)		
	Environment		
	Quiet	TV-news	Music
Close-talking	94.4	91.7	91.9
NAM	93.8	93.2	92.9

a simulation experiment using 24 nonaudible murmur utterances and 30 audible speech utterances. Figure 22 shows the histogram of the duration normalized scores of the two GMMs, when the input signal is audible speech. As can be seen, in all the cases the score of the GMM corresponding to normal speech (S_N) is higher than the score of the GMM corresponding to nonaudible murmur speech. Therefore, based on these scores the HMMs set is selected correctly. Figure 23 shows the histogram of the GMM scores when the input signal is nonaudible murmur. The figure shows that the scores of a nonaudible murmur GMM are higher, and therefore the correct HMMs set is selected in this case, too. The system achieved a 92.1% word accuracy on average, which is a very promising result. A single recognizer using the same NAM models and the same NAM test utterances achieved a 90.4% word accuracy. Using the same normal-speech test utterances and the same NAM models, the word accuracy was only 4.7%. Although the system shows high performance, the delay necessary for the GMM matching is a disadvantage.

6.2. Using parallel speech recognizers

To overcome the problem of the delay, we introduce another method based on parallel speech recognizers. Two recognizers using different HMMs (audible speech, nonaudible murmur) operate in parallel providing two hypotheses with their scores. The system selects the hypothesis with the higher score as the correct recognition result. Figure 24 shows the block diagram of the system. Using the same test set as in the previous section, the system achieved a 92.1% word accuracy in this case, too. The disadvantage of this method is the higher complexity due to the use of two recognizers.

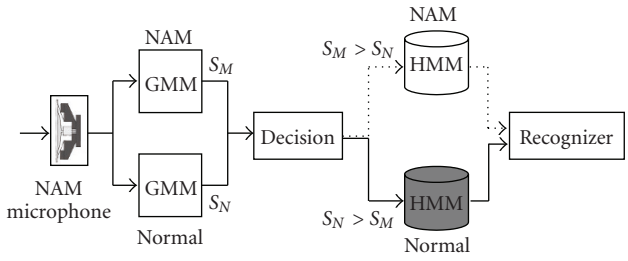

Figure 21: GMM-based discrimination.

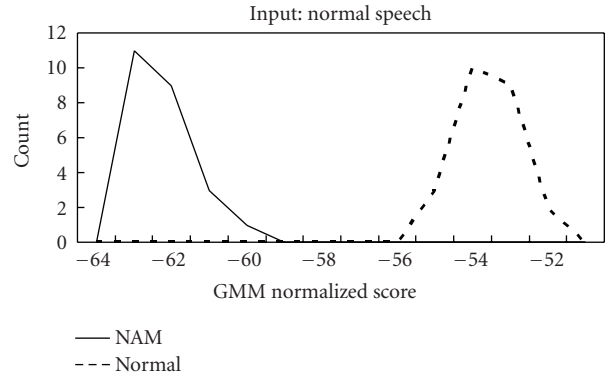

Figure 22: GMM normalized scores—input normal speech.

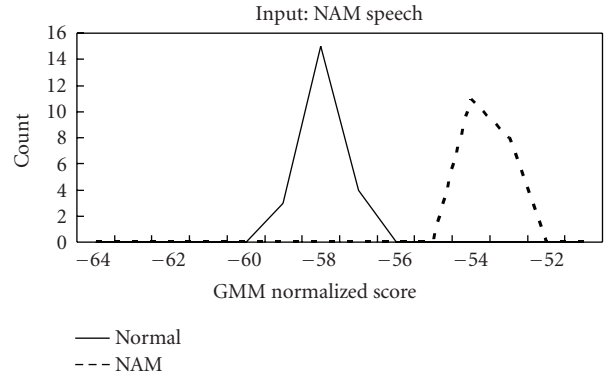

Figure 23: GMM normalized scores—input nonaudible murmur.

6.3. Using a combined HMM set

In this experiment, only an HMMs set was used trained with nonaudible murmur data and audible speech data recorded using a NAM microphone. For MLLR adaptation we used the

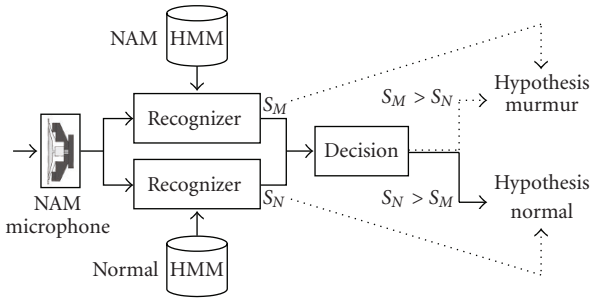

Figure 24: Parallel recognizers-based recognition.

same data as in Sections 6.1 and 6.2. Using this approach, we achieved a 91.4% word accuracy on average. The results show that this is a very effective approach and does not require additional sources. On the other hand, the performance of this approach depends on the ratio of the two different training data used to train the combined HMM set. Our experience showed that a larger nonaudible murmur training database is required.

7. CONCLUSIONS

In this paper, we presented nonaudible murmur recognition in clean and noisy environments using NAM microphones. A NAM microphone is a special acoustic device attached behind the talker's ear, which can capture very quietly uttered speech. Nonaudible murmur recognition can be used when privacy in human-machine communication is desired. Since nonaudible murmur is captured directly from the body, it is less sensitive to environmental noises. To show this, we carried out experiments using simulated and real noisy data. Using simulated noisy data at 50 dBA and 60 dBA noise levels, the nonaudible murmur recognition performance was almost equal to that of the clean case. Using, however, data recorded in noisy environments, the performance decreased. To investigate the possible reasons for this, we studied the role of the Lombard effect in nonaudible murmur recognition and we carried out an experiment using Lombard data. The results showed that the Lombard reflex has a negative impact effect on nonaudible murmur recognition. Due to the speech production modifications, the nonaudible murmur characteristics under Lombard conditions are changed and show a higher similarity to normal speech. Due to this fact, a mismatch appears between the training and testing conditions and the performance decreases. We also proposed a method based on multilevel Lombard HMMs set to recognize arbitrary Lombard nonaudible murmur utterances.

In this paper, we also reported audible speech recognition using NAM microphone showing the effectiveness of a NAM microphone in normal speech recognition. To make a nonaudible murmur-based system more flexible and general, we introduced three approaches to integrate normal speech recognition and nonaudible murmur recognition using a NAM microphone with promising results.

In this paper, we reported speaker-dependent applications of NAM microphones. As future work, we plan to investigate nonaudible murmur recognition in the speaker-independent domain. Currently, collection of nonaudible murmur database from several speakers is in progress. Also, sampling the NAM data at 8 kHz seems to be more appropriate due to the limited high frequency band of NAM.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Yoshitaka Nakajima for providing the NAM microphones and also all the members of the Speech and Acoustics Processing Laboratory for their collaboration in collecting nonaudible murmur data.

REFERENCES

[1] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 5, pp. 708–711, Hong Kong, April 2003.

[2] Y. Zheng, Z. Liu, Z. Zhang, et al., "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '03)*, pp. 249–254, St. Thomas, Virgin Islands, USA, November-December 2003.

[3] Z. Liu, A. Subramanya, Z. Zhang, J. Droppo, and A. Acero, "Leakage model and teeth clack removal for air- and bone-conductive integrated microphones," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 1, pp. 1093–1096, Philadelphia, Pa, USA, March 2005.

[4] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Processing Letters*, vol. 10, no. 3, pp. 72–74, 2003.

[5] O. M. Strand, T. Holter, A. Egeberg, and S. Stensby, "On the feasibility of ASR in extreme noise using the PARAT earplug communication terminal," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '03)*, pp. 315–320, St. Thomas, Virgin Islands, USA, November-December 2003.

[6] S.-C. Jou, T. Schultz, and A. Waibel, "Adaptation for soft whisper recognition using a throat microphone," in *Proceedings of International Conference on Speech and Language Processing (ICSLP '04)*, Jeju Island, Korea, October 2004.

[7] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition," in *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, pp. 2601–2604, Geneva, Switzerland, September 2003.

[8] A. Lee, T. Kawahara, K. Takeda, and K. Shikano, "A new phonetic tied-mixture model for efficient decoding," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, vol. 3, pp. 1269–1272, Istanbul, Turkey, June 2000.

[9] 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," in *Proceedings of IEEE Workshop on Automatic*

Speech Recognition and Understanding (ASRU '03), pp. 73–76, St. Thomas, Virgin Islands, USA, November-December 2003.

[10] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano, "Non-audible murmur (NAM) recognition using a stethoscopic NAM microphone," in *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech '04 - ICSLP)*, pp. 1469–1472, Jeju Island, Korea, October 2004.

[11] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano, "Applications of NAM microphones in speech recognition for privacy in human-machine communication," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05 - EUROSPEECH)*, pp. 3041–3044, Lisboa, Portugal, September 2005.

[12] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Remodeling of the sensor for non-audible murmur (NAM)," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05 - EUROSPEECH)*, pp. 389–392, Lisboa, Portugal, September 2005.

[13] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano, "Investigating the role of the Lombard reflex in non-audible murmur (NAM) recognition," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05 - EUROSPEECH)*, pp. 2649–2652, Lisboa, Portugal, September 2005.

[14] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano, "Audible (normal) speech and inaudible murmur recognition using NAM microphone," in *Proceedings of the 7th European Signal Processing Conference (EUSIPCO '04)*, pp. 329–332, Vienna, Austria, September 2004.

[15] T. Kawahara, A. Lee, T. Kobayashi, et al., "Free software toolkit for Japanese large vocabulary continuous speech recognition," in *Proceedings of 6th International Conference on Spoken Language Processing (ICSLP '00)*, pp. IV-476–IV-479, Beijing, China, October 2000.

[16] K. Itou, M. Yamamoto, K. Takeda, et al., "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *The Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[17] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[18] C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 806–814, 1991.

[19] P. C. Woodland, D. Pye, and M. J. F. Gales, "Iterative unsupervised adaptation using maximum likelihood linear regression," in *Proceedings of the 4th International Conference on Spoken Language (ICSLP '96)*, vol. 2, pp. 1133–1136, Philadelphia, Pa, USA, October 1996.

[20] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[21] A. Wakao, K. Takeda, and F. Itakura, "Variability of Lombard effects under different noise conditions," in *Proceedings of the 4th International Conference on Spoken Language (ICSLP '96)*, vol. 4, pp. 2009–2012, Philadelphia, Pa, USA, October 1996.

[22] J. H. L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 598–614, 1994.

[23] B. A. Hanson and T. H. Applebaum, "Robust speaker-independent word recognition using static, dynamicand acceleration features: experiments with Lombard and noisy speech," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP '90)*, vol. 2, pp. 857–860, Albuquerque, NM, USA, April 1990.

[24] R. Ruiz, E. Absil, B. Harmegnies, C. Legros, and D. Poch, "Time- and spectrum-related variabilities in stressed speech under laboratory and real conditions," *Speech Communication*, vol. 20, no. 1-2, pp. 111–129, 1996.

Panikos Heracleous was born in Paphos, Cyprus, on May 22, 1966. He received the M.S. degree in electrical engineering from the Technical University of Budapest, Hungary, in 1992, and the Dr. Eng. degree from the Nara Institute of Science and Technology, Japan, in 2002. In 2001, he joined KDDI R&D Labs as a Research Engineer in telephone speech recognition field. In 2003, he joined the Speech and Acoustics Processing Laboratory, Nara Institute of Science and Technology as a CEO Postdoctoral Research Fellow. During the period from October 2005 to January 2006, he was an Assistant Professor at Nara Institute of Science and Technology. He is currently an Assistant Professor at University of Cyprus. His research interests include signal processing, microphone arrays, automatic speech recognition, and unvoiced speech recognition. He is a Member of ISCA, IEEE, IEICE, and the Acoustical Society of Japan.

Tomomi Kaino received the B.S. degree in information and computer science from Nara Woman's University in 2004, and the M.S. degree in information science from Nara Institute of Science and Technology in 2006. She had been studying in body-transmitted speech recognition in multi-speaking styles including nonaudible murmur (NAM) in her Master's course. She is now working with Sanyo Electric Co., Ltd.

Hiroshi Saruwatari was born in Nagoya, Japan, on July 27, 1967. He received the B.E., M.E., and Ph.D. degrees in electrical engineering from Nagoya University, Nagoya, Japan, in 1991, 1993, and 2000, respectively. He joined Intelligent Systems Laboratory, Secom Co., Ltd., Mitaka, Tokyo, Japan, in 1993, where he engaged in the research and development of the ultrasonic array system for the acoustic imaging. He is currently an Associate Professor of Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include array signal processing, blind source separation, and sound field reproduction. He received the Paper Awards from IEICE in 2001 and 2006. He is a Member of the IEEE, the VR Society of Japan, the IEICE, and the Acoustical Society of Japan.

Kiyohiro Shikano received the B.S., M.S., and Ph.D. degrees in electrical engineering from Nagoya University in 1970, 1972, and 1980, respectively. He is currently a Professor at Nara Institute of Science and Technology (NAIST), where he is directing Speech and Acoustics Laboratory. From 1972, he had been working at NTT Laboratories, where he had been engaged in speech recognition research. During 1990–1993, he was the Executive Research Scientist at NTT Human Interface Laboratories, where he supervised the research of speech recognition and speech coding. During 1986–1990, he was the Head of Speech Processing Department at ATR Interpreting Telephony Research Laboratories, where he was directing speech recognition and speech synthesis research. During 1984–1986, he was a Visiting Scientist at Carnegie Mellon University. He received the Institute of Electronics, Information and Communication Engineers of Japan (IEICE) Yonezawa Prize in 1975, IEEE Signal Processing Society 1990 Senior Award in 1991, the Technical Development Award from Acoustical Society of Japan (ASJ) in 1994, Information Processing Society of Japan (IPSJ) Yamashita SIG Research Award in 2000, Paper Award from the Virtual Reality Society of Japan in 2001, IEICE Paper Award in 2005 and 2006, and IEICE Inose Best Paper Award in 2005. He is a Fellow Member of IEICE and IPSJ. He is a member of ASJ, Japan VR Society, IEEE, and International Speech Communication Association.