AUDIBLE (NORMAL) SPEECH AND INAUDIBLE MURMUR RECOGNITION USING NAM MICROPHONE

Panikos Heracleous, Yoshitaka Nakajima, Akinobu Lee, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science
Nara Institute of Science and Technology, Japan
8916-5 Takayama-cho Ikoma-shi Nara 630-0192, Japan
e-mail: {panikos,yoshi-n,ri,sawatari,shikano}@is.aist-nara.ac.jp

ABSTRACT

In this paper, we present audible (normal) speech and inaudible murmur hidden Markov models based automatic speech recognition using NAM microphone. The NAM (Non-Audible Murmur) microphone is a special device, which can be used for capturing inaudible murmur speech. The device is based on the stethoscope, which is used in medical science. By attaching the NAM microphone behind the talker's ear, we can receive very quietly uttered speech and perform automatic speech recognition in a conventional way. Privacy, robustness to environmental, and a useful tool for sound-impaired people noise belong to the advantages of the NAM microphone. Using adaptation techniques, we created hidden Markov models for inaudible speech and we performed automatic speech recognition. The achieved results are very promising, and prove the effectiveness of NAM microphone In this paper, we also introduce our work for recognizing normal speech using NAM microphone. The idea is to take advantage of noise robustness of NAM microphone. In our experiments, we achieved a 93.8% word accuracy in clean environment, and a 93.1% word accuracy in noisy environment. In this paper, we also introduce two techniques to intergrate inaudible murmur and audible speech recognition using NAM microphone. In both cases, we achieved a 92.1% word accuracy on average, which is a very promising result.

1. INTRODUCTION

The NAM (Non-Audible Murmur) microphone [1] is a special, new device able to capture inaudible murmur speech (NAM speech), which cannot be heared by listeners near the talker. The device is based on the stethoscope, which is used in medical science. By attaching the NAM microphone behind the talker's ear, we can receive very quietly uttered speech, and perform automatic speech recognition in a convectional way. Privacy, robustness to environmental noise, and a useful tool for sound-impaired people are the advantages of NAM microphone, when it is applied in a speech recognition system. Figure 1 shows the attachment of NAM microphone to the talker. The optimal position of the attachment was determined experimentally. Although the NAM signal is of poor quality, the signal envelope is similar to that of normal speech, and therefore speech recognition is possible.

In previous work [2], we introduced the experiments and results for recognition of inaudible murmur. Using speech received by NAM microphone, we created hidden Markov models (HMM) for inaudible speech recognition. Instead of creating speaker-independent HMMs, or speaker-dependent

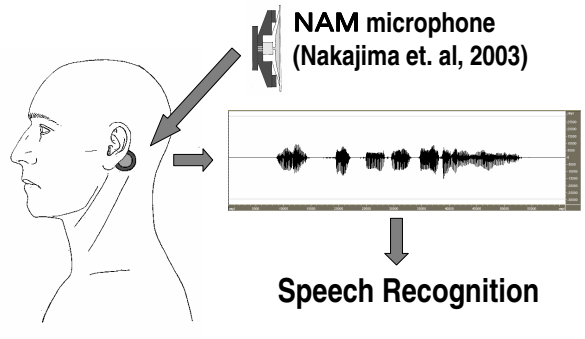

Figure 1: NAM microphone attached to the talker

HMMs trained with a large amount of data, we used adaptation techniques to create acoustic models for inaudible murmur recognition. Adaptation is a widely used technique for creating speaker, or environment specific acoustic models, when limited training data are available. The maximum likelihood linear regression adaptation (MLLR) technique [3] was selected in our work.

We carried out experiments using two kind of initial HMMs. In the first case, we used normal-speech HMMs as initial models. In the second case, we extended our work by collecting NAM data from several speaker to train NAM initial models. Since the NAM speech characteristics are different from normal speech characteristics, a modified version of the MLLR was used in our work. More specifically, due to the big difference between initial models and adaptation data, the conventional single-iteration MLLR is not effective in NAM recognition. The iterative MLLR appears to be more effective, and results show that it provides higher performance. Our proposed method is similar to that proposed by Woodland et. al,[4], but the object is different. However, we try to increase the performance of MLLR by increasing the number of iterations and using the same adaptation data at each pass, based on the fact the MLLR is based on Expectation Maximization (EM) method. Figure 2 show the proposed method. The initial models are adapted using the MLLR technique and a small amount of adaptation utterances. As a result, intermediate models are created. The intermediate adapted models are re-adapted using the same adaptation data, and this procedure is continued until no further improvement is obtained.

Figure 3 shows the results of inaudible murmur recognition. In this experiment, the recognition engine is Julius 20k vocabulary Japanese Dictation Toolkit [6]. For test-

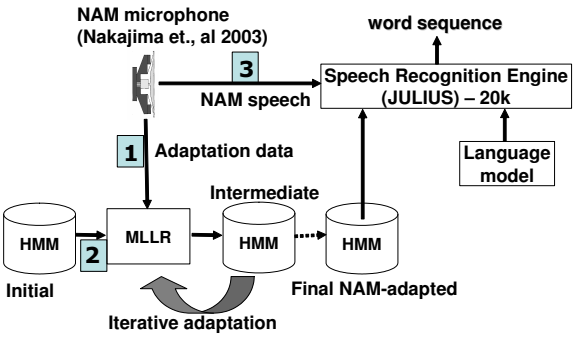

Figure 2: Iterative MLLR for NAM recognition

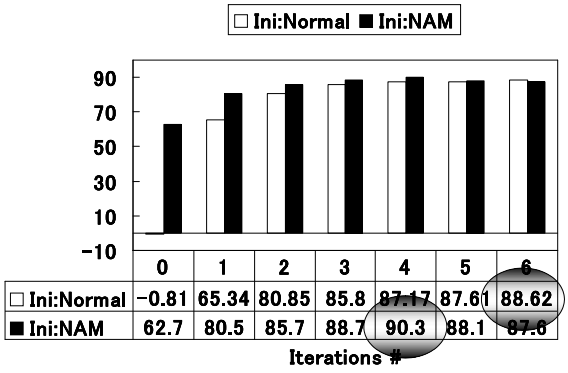

Figure 3: Word accuracy of inaudible murmur recognition

ing, 72 utterances were used recorded under several environments (quiet, music, TV-news). The achieved results show the effectiveness of the iterative MLLR. As can be seen, with single iteration MLLR the performance of the system is very low. By increasing the number of iterations, the performance is drastically increased. More specifically, using normal-speech initial HMMs we achieved after 6 iterations a 88.62% word accuracy. Using NAM-speech initial HMMs we achieved after 4 iterations a 90.3% word accuracy. Results show also the importance of the initial HMMs. In the case of NAM-speech initial models, the initial performance is 62.7% (-0.81% when using normal-speech initial models). Therefore, with less number of iterations we achieved higher performance.

2. AUDIBLE (NORMAL) SPEECH RECOGNITION USING NAM MICROPHONE

The obtained results show the effectiveness of NAM microphone in inaudible murmur recognition. Using NAM microphone and a small amount of adaptation data, we recognized speech uttered very quietly with very high accuracy. Therefore, NAM microphone 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 able to recognize audible (normal) speech, too.

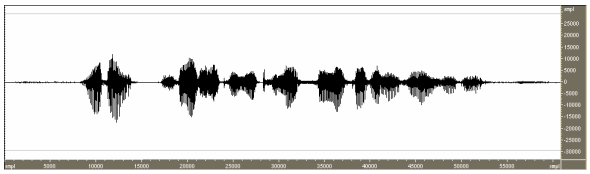

Figure 4: Normal speech waveform - Close-talking microphone

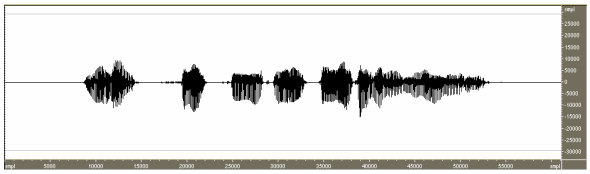

Figure 5: Normal speech waveform - NAM microphone

In this paper, we focus on this problem. First, we prove that NAM microphone can be used very effectively for audible (normal) speech recognition, taking also advantage of its robustness against noise. In the following, we implement two baseline approaches to integrate audible speech and inaudible murmur recognition. The first approach is a GMM (Gaussian Mixture Model) based discrimination, and the second one is based on parallel speech recognizers. Experimental results show very high performance for both methods.

Figure 4 shows the waveform of a normal-speech signal received by a close-talking microphone. Figure 5 shows the same signal received by a NAM microphone. The two signals are synchronized, due to a two-channels recording. The figures show the high similarity between the two signals. Figures 6 and 7 show the spectrograms of the received speech signals. As can be seen, the signal received by the NAM microphone has limited frequency band, compared to signal received by the close-talking microphone. As a result, the quality of the signal received by the NAM microphone is lower. For speech recognition, however, the similarity in the signal envelope is sufficient. The different frequency characteristics of the two signals require different approach for speech recognition. More specifically, the acoustic models used to recognize normal 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 create a new acoustic models set.

The HMM set for recognition of audible speech received by NAM microphone is created using iterative MLLR. A 128-class regression tree, 350 adaptation utterances, and 4 iterations are used. The initial HMMs used in these experiments are Phonetic Tied Mixture (PTM) models with 3000 states [7]. The models are trained using the speech corpus collected by the Acoustical Society of Japan [5]. For evaluation, 72 NAM utterances, recorded under several conditions (quiet, background music, TV-news) are used. The speech recognition engine is the Julius 20k vocabulary Japanese Dictation Toolkit. For comparison, we created HMM for recognition of normal speech received by a close-talking microphone. Single-iteration MLLR with 32-class regression tree, and 100 adaptation utterances is used. Table 1 shows the system specifications.

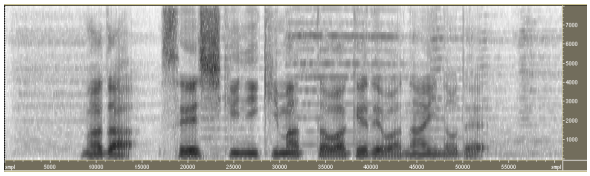

Figure 6: Normal speech spectrogram - Close-talking microphone

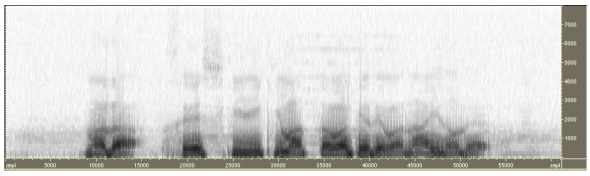

Figure 7: Normal speech spectrogram - NAM microphone

Table 2 shows the achieved results. As can be seen, in quiet environment the speech received by NAM microphone is recognized with slightly lower accuracy. The reason is the information lost during body transmission. In the case, however, when there is background noise (music, TV-news) the recognition of audible speech received by NAM microphone shows higher performance. Although under noisy environments the performance decreases, we observe that the decreases is 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 are very promising and prove the effectiveness of NAM microphone for audible speech recognition, too. Especially, in noisy environments this is a very important advantage.

3. INTEGRATED AUDIBLE AND INAUDIBLE MURMUR RECOGNITION

A challenging topic is to integrate audible (normal) and inaudible murmur recognition. In the previous sections, we showed the effectiveness of NAM microphone in inaudible murmur and audible (normal) speech recognition. A recognition system, which combines recognition of the two kind of speeches using 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. Moreover, in noisy environments a NAM microphone based system shows significant robustness against environmental noise. In this section, we introduce two techniques to integrate inaudible murmur and audible speech recognition. Both approaches are based on case-dependent HMMs created using iterative MLLR.

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 ΔMFCCs 1-order ΔE
HMM	PTM , 3000 states
Training data	JNAS database
Test data	72 NAM utterances

Table 2: Recognition rates for audible speech

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

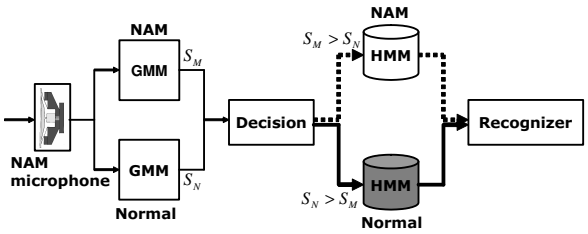

Figure 8: GMM based discrimination

3.1 Gaussian Mixture Models (GMM) based discrimination

The first approach is based on GMM based discrimination. Two GMM (one-emitting state HMM) were trained using audible speech and inaudible murmur received by NAM microphone, respectively. The transcriptions of the uttered speech were merged to form only one model. Figure 8 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 HMM sets used in this experiment are the same as in the experiments described in the previous sections.

To evaluate the performance of the method, we carried out a simulation experiment using 24 inaudible murmur utterances and 30 audible speech utterances. Figure 9 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 GMM corresponding to inaudible murmur speech. Therefore, based on these scores the HMM set is selected correctly. Figure 10 shows the histogram of the GMM scores, when the input signal is inaudible murmur. The figure shows, that the scores of inaudible murmur GMM are higher, and therefore the correct HMM set is selected in this case, too. The system achieved a 92.1% word accuracy on average, which is a very promising result. Although the system shows high performance, the delay necessary for the GMM matching is a disadvantage.

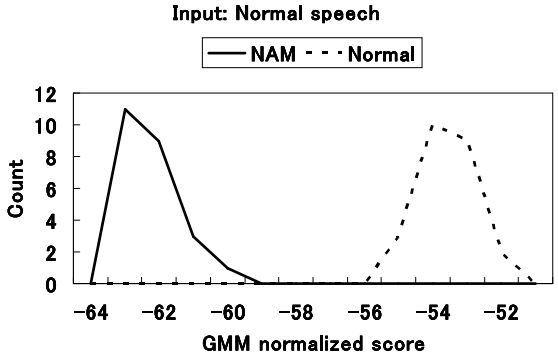

Figure 9: GMM normalized scores - Input normal speech

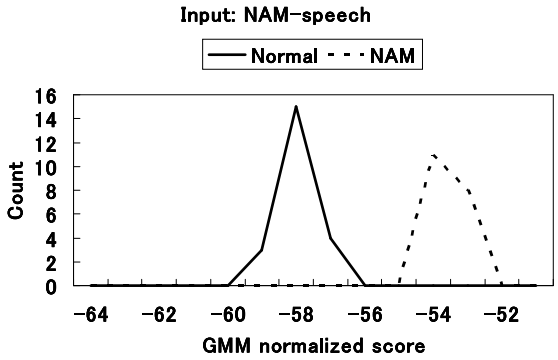

Figure 10: GMM normalized scores - Input inaudible murmur

3.2 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, inaudible 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 11 shows the block diagram of the system. Table 3 show the comparisons of the two scores. As can be seen, the appropriate hypothesis is correctly selected. 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 two recognizers.

4. CONCLUSION - FUTURE WORK

In this paper, we introduced results of experiments, which prove the effectiveness of NAM microphone in audible (normal) speech recognition, too. Using a NAM microphone as input device for normal speech recognition, we achieved in clean environment a 93.8% word accuracy, and in noisy enviroment a 93.1% word accuracy on average. We also implement two basic approaches to intergate audible speech and inaudible murmur recognition based on NAM microphone. A GMM based discrimination and a parallel speech recognizers based method were presented. Both methods achieved a 92.1% word accuracy on average, which is a very promising result. As future work, we plan to use NAM microphone to recognize audible speech and inaudible murmur in more noisy environments. The experiments described in this paper, use speaker-dependent HMMs trained using MLLR adaptation. Currently, we are collecting speech data using NAM microphone to train speaker-independent HMMs.

Table 3: Parallel speech recognition based integration

	DECISIONS	
TEST SET	$S_M > S_N$	$S_N > S_M$
24 NAM	24	0
30 Normal	0	30

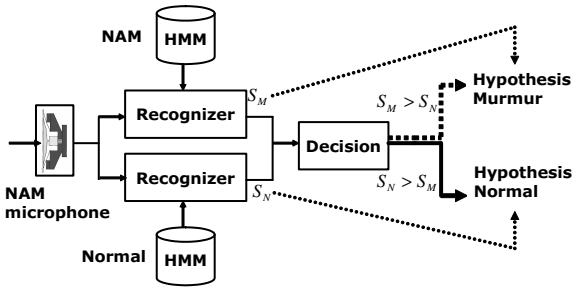

Figure 11: Parallel recognizers based recognition

REFERENCES

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

[2] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, K. Shikano " Accurate Hidden Markov Models for Non-Audible Murmur (NAM) Recognition Based on Iterative Supervised Adaptation", *Proceedings of ASRU*, pp. 73–76, 2003.

[3] C. J. Leggetter, C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", *Computer Speech and Language*, Vol. 9, pp. 171–185, 1995.

[4] P.C Woodland, D. Pye, M.J.F. Gales, "Iterative Unsupervised Adaptation Using Maximum Likelihood Linear Regression", *Proceedings of ICSLP*, pp. 1133–1136, 1996.

[5] K. Itou et al., "JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research", *The Journal of Acoustical Society of Japan (E)*, Vol. 20, pp. 199–206, 1999.

[6] T. Kawahara et al., "Free Software Toolkit for Japanese Large Vocabulary Continuous Speech Recognition", *Proceedings of ICSLP*, pp. IV-476–479, 2000.

[7] A. Lee, T. Kawahara, K. Takeda, K. Shikano, "A New Phonetic Tied Mixture Model for Efficient Decoding", *Proceedings of ICASSP*, pp. 1269–1272, 2000.