# ACCURATE HIDDEN MARKOV MODELS FOR NON- AUDIBLE MURMUR (NAM) RECOGNITION BASED ON ITERATIVE SUPERVISED ADAPTATION

*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 previous works, we introduced a special device (Non-Audible Murmur (NAM) microphone) able to detect very quietly uttered speech (murmur), which cannot be heard by listeners near the talker. Experimental results showed the efficiency of the device in NAM recognition. Using normal-speech monophone hidden Markov models (HMM) retrained with NAM data from a specific speaker, we could recognize NAM with high accuracy. Although the results were very promising, a serious problem is the HMM retraining, which requires a large amount of training data. In this paper, we introduce a new method for NAM recognition, which requires only a small amount of NAM data for training. The proposed method is based on supervised adaptation. The main difference from other adaptation approaches lies in the fact that instead of single-iteration adaptation, we use iterative adaptation (iterative supervised MLLR). Experiments prove the efficiency of the proposed method. Using normal-speech clean initial models and only 350 adaptation NAM utterances, we achieved a recognition accuracy of 88.62%, which is a very promising result. Therefore, with a small amount of adaptation data, we were able to create accurate individual HMMs. We also introduce results of experiments, which show the effects of the number of iterations, the amount of adaptation data, and the regression tree classes.

## 1. INTRODUCTION

Non-Audible Murmur (NAM) is speech uttered very quietly, which cannot be heard by listeners near the speaker. Using a special device (NAM microphone), which is attached directly to the head, we can receive NAM signals and perform automatic speech recognition [1]. The NAM microphone is based on a device used in medical science (stethoscope) and can detect very quietly uttered speech (murmur). Although the received speech signal is of poor quality, the envelope of the NAM signal is similar to that of normal speech, and therefore, speech recognition is possible.

There are three advantages of NAM recognition.

- Privacy
For practical speech recognition systems (e.g. telephone speech recognition system) privacy plays an important role. However, users should be able to communicate with the speech recognition engine without others hearing their conversations. The conventional speech recognition systems, however, are based on normal speech recognition and cannot effectively provide privacy. In contrast, NAM recognition allows users to use a speech recognition device with privacy.

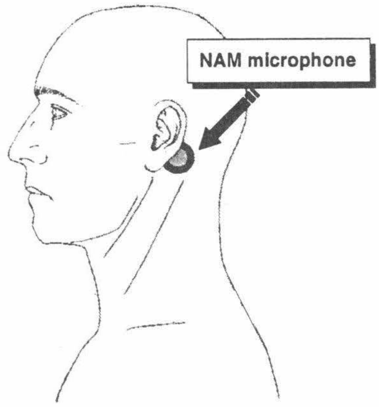


**Fig. 1**. NAM microphone attached to the talker

- Robustness to noise
A serious problem in speech recognition is environmental noise. In normal speech recognition, the speech signal is distorted by noise and the performance of the system is degraded. With the NAM microphone, the signal is received directly from the body and therefore is more robust against environmental noise.
- Useful tool for speech (sound)-impaired people
For people suffering from physical difficulties in speech (speech or sound impairments), NAM recognition can provide a useful tool for communicating with a machine. Moreover, speech recognition can be combined with speech synthesis, allowing speech-impaired people to communicate in a natural manner.

Figure 1 shows the attachment of the NAM to the head.

## 2. PROPOSED ADAPTATION METHOD FOR NAM RECOGNITION

The recognition of NAM requires accurate hidden Markov models (HMMs). The ideal procedure is to train speaker-independent NAM HMMs. However, this requires a large amount of training data which are very difficult to collect in a short time. On the other

0-7803-7980-2/03/$17.00 © 2003 IEEE

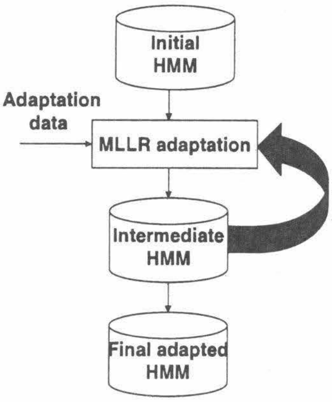


**Fig. 2.** Iterative MLLR for NAM recognition

**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,<br>12-order ΔMFCCs<br>1-order ΔE |
| HMM | PTM , 3000 states<br>3 states, 16 distributions<br>monophones |
| Training data | JNAS database |

hand, the use of a NAM-based recognition engine is strictly individual, and therefore, speaker-dependent models are also efficient. Three possible ways of training NAM HMMs are as follows:

- Train speaker- and gender-dependent models. For this purpose, however, a large amount of training data, which must be collected, is required.
- Train speaker-dependent models for a specific user. However, for accurate HMMs a large amount of training data is again necessary.
- Use available HMMs as initial models, and perform adaptation to a specific speaker's characteristics. This method requires only a small amount of adaptation data, and can be easily applied. The maximum likelihood linear regression (MLLR) adaptation technique [2] was selected in our work.

Since the NAM data are different from normal speech data, a modified version of the MLLR was used in this work. More specifically, due to the HMM distance between initial and adapted models, 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. 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. In our proposed method, similar components are clustered together using a regression tree. Acoustically similar components are transformed in the same way.

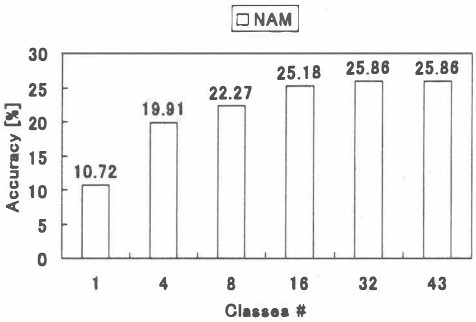


**Fig. 3.** Effect of classes of regression tree - monophone initial models

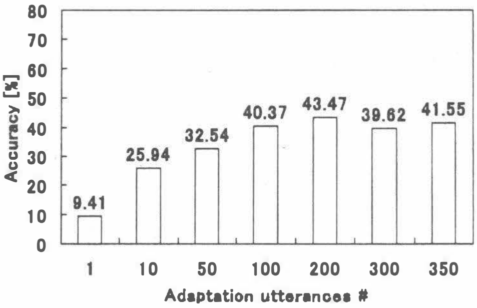


**Fig. 4.** Effect of amount of adaptation data - monophone initial models

## 3. EXPERIMENTS AND RESULTS

### 3.1. Using monophone initial models

In these experiments, monophone initial models are used. Forty-three Japanese monophone models are trained with the speech corpus collected by the Acoustical Society of Japan [3]. The topology of the HMMs are 3 states left to right with no skip. Sixteen Gaussian mixtures per state are used. For evaluation, 72 NAM utterances, recorded under several conditions (quite, background music, TV-news) are used. The speech recognition engine is the Julius 20k vocabulary Japanese Dictation Toolkit [4].

Figure 3 shows the results obtained using several class regression trees. In this experiment, 25 adaptation utterances were used and the MLLR was performed in a single iteration. Results show that by increasing the number of classes, improvement was achieved. In the following, a 32-class regression tree is used. Figure 4 shows the performance for various amounts of adaptation data. Using a 32-class regression tree and 200 adaptation utterances, a recognition accuracy of 43.47% was achieved.

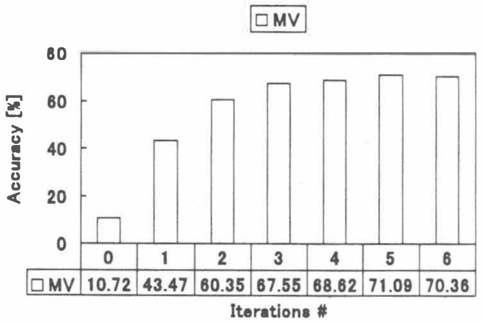


**Fig. 5.** Effect of number of iterations - monophone initial models

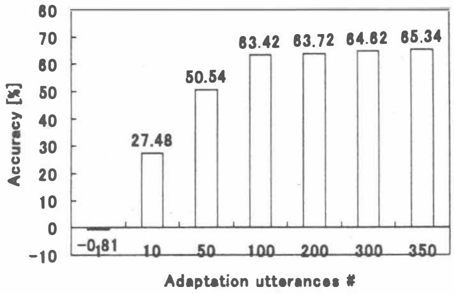


**Fig. 7.** Effect of amount of adaptation data - PTM initial models

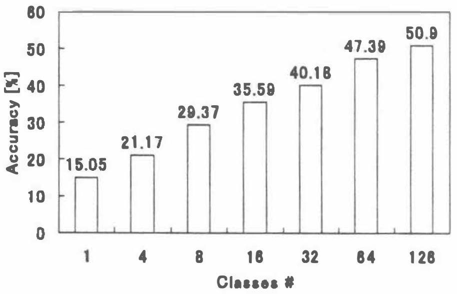


**Fig. 6.** Effect of classes of regression tree - PTM initial models

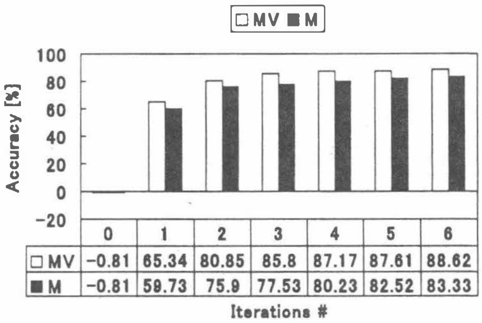


**Fig. 8.** Effect of number of iterations of MLLR - PTM models - NAM

Figure 5 shows the results when iterative MLLR was performed. In this experiment, 200 adaptation utterances and a 32-class regression tree were used. As can been seen, after 5 iterations, the recognition accuracy was significantly increased to 71.09%.

### 3.2. Using Phonetic Tied Mixture (PTM) initial models

The HMMs used in these experiments are Phonetic Tied Mixture (PTM) models with 3000 states [5]. The models are trained using the speech corpus collected by the Acoustical Society of Japan. For evaluation, 72 NAM utterances, recorded under several conditions (quite, background music, TV-news) are used. The speech recognition engine is the Julius 20k vocabulary Japanese Dictation Toolkit.

Figure 6 shows the results obtained using several class regression trees. In this experiment, 25 adaptation utterances were used and the MLLR was performed in a single iteration. As can be seen, by increasing the number of classes to 128, significant improvement was achieved.

Figure 7 shows the performance for various amounts of adaptation data. In this experiment, a 128-class regression tree was used and single-iteration MLLR was performed. Figure 7 shows that when using 350 adaptation utterances, a recognition accuracy of 65.34% was obtained.

Figure 8 shows the performance when iterative MLLR was performed. Without adaptation (0 iterations), the recognition accuracy using the initial clean PTM models is -0.81%. After 6 iterations, the recognition accuracy increases significantly to a promising 88.62%. Figure 8 shows also that higher performance is achieved when both means and variances (MV) are transformed, instead of only means (M).

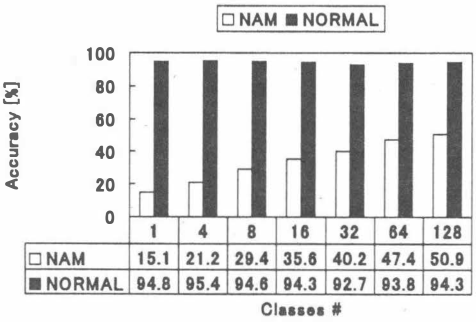
The results achieved in these experiments also show, that using PTM, instead of monophone models, significant improvement was achieved. More specifically, using monophone models, a recognition accuracy of 71.09% was obtained. Using, however, PTM models, a recognition accuracy of 88.62% was achieved.

### 3.3. Comparison between NAM and normal speech recognition

In order to evaluate the performance using normal speech, we carried out experiments with the same adaptation and evaluation data, recorded using a close-talking microphone in a clean environment.

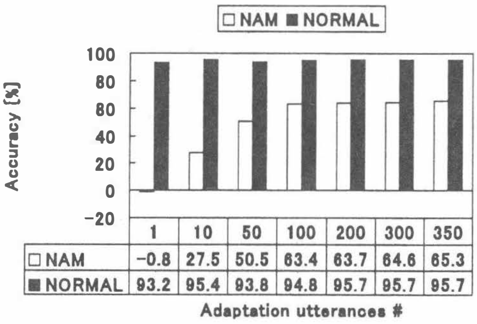
Figure 9 shows the results obtained using several class regression trees. In this experiment, 25 adaptation utterances were used and the MLLR was performed in a single iteration. As the results show, in the case of normal speech, the recognition accuracy does not change significantly. We can also observe that the initial performance (1 adaptation utterance) is very high (94.8%).

Figure 10 show the results of using various amounts of adaptation data. Using normal speech, a recognition accuracy of 95.7% was obtained. In this experiment, a 32-class regression tree was



| | 1 | 4 | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|---|---|
| □ NAM | 15.1 | 21.2 | 29.4 | 35.6 | 40.2 | 47.4 | 50.9 |
| ■ NORMAL | 94.8 | 95.4 | 94.6 | 94.3 | 92.7 | 93.8 | 94.3 |

**Fig. 9**. Effect of classes of regression tree - PTM initial models



| | 1 | 10 | 50 | 100 | 200 | 300 | 350 |
|---|---|---|---|---|---|---|---|
| □ NAM | −0.8 | 27.5 | 50.5 | 63.4 | 63.7 | 64.6 | 65.3 |
| ■ NORMAL | 93.2 | 95.4 | 93.8 | 94.8 | 95.7 | 95.7 | 95.7 |

**Fig. 10**. Effect of amount of adaptation data - PTM initial models

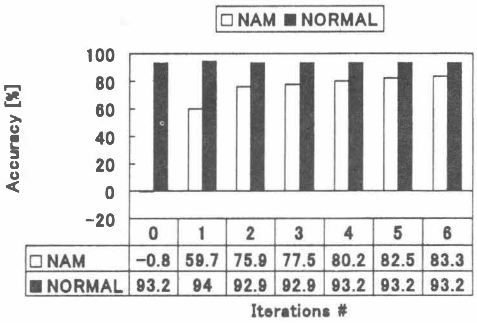
used and the MLLR was performed in a single iteration.

Figures 11 and 12 show the results of using iterative MLLR. In the case of normal speech, the iterative MLLR is not effective. In fact, on increasing the number of iterations, the performance decreases. In the case of NAM, however, using iterative MLLR, the performance is increased significantly. Although the performance is higher when using normal speech (due to the initial performance), the best results of the two cases might be considered to be comparable.

## 4. CONCLUSION - FUTURE WORK

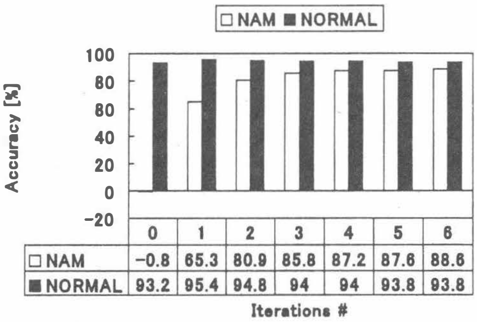
In this paper, we introduced a new supervised adaptation technique for NAM recognition. Experiments confirmed the efficiency of the proposed iterative adaptation. Using the proposed method and a small amount of adaptation data, we achieved a recognition accuracy of 88.62%, which is a very promising result. Therefore, using a minimal amount of data, we were able to create individual NAM models. However, problems still remain, some examples of which are as follows:

- Initial performance
- Practical body attachment
- Sensitivity to internal noises
- Non-natural communication using NAM



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| □ NAM | −0.8 | 59.7 | 75.9 | 77.5 | 80.2 | 82.5 | 83.3 |
| ■ NORMAL | 93.2 | 94 | 92.9 | 92.9 | 93.2 | 93.2 | 93.2 |

**Fig. 11**. Effect of number of iterations of MLLR - PTM initial models - M



| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| □ NAM | −0.8 | 65.3 | 80.9 | 85.8 | 87.2 | 87.6 | 88.6 |
| ■ NORMAL | 93.2 | 95.4 | 94.8 | 94 | 94 | 93.8 | 93.8 |

**Fig. 12**. Effect of number of iterations of MLLR - PTM initial models - MV

As future work, we plan to investigate the problems described above.

## 5. REFERENCES

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

[2] 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.

[3] 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.

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

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