A tissue-conductive acoustic sensor applied in speech recognition for privacy

Panikos Heracleous, Yoshitaka Nakajima, Hiroshi Saruwatari and Kiyohiro Shikano

Nara Institute of Science and Technology, Japan
{panikos,yoshi-n,sawatari,shikano}@is.naist.jp

Abstract

In this paper, we present the Non-Audible Murmur (NAM) microphones focusing on their applications in automatic speech recognition. A NAM microphone is a special acoustic sensor attached behind the talker's ear and able to capture very quietly uttered speech (non-audible murmur) through body tissue. Previously, we reported experimental results for non-audible murmur recognition using a Stethoscope microphone in a clean environment. In this paper, we also present a more advanced NAM microphone, the so-called Silicon NAM microphone. Using a small amount of training data and adaptation approaches, we achieved a 93.9% word accuracy for a 20k vocabulary dictation task. Therefore, in situations when privacy in human-machine communication is preferable, NAM microphone can be very effectively applied for automatic recognition of speech inaudible to other listeners near the talker. Because of the nature of non-audible murmur (e.g., privacy) investigation of the behavior of NAM microphones in noisy environments is of high importance. To do this, we also conducted experiments in real and simulated noisy environments. Although, using simulated noisy data the NAM microphones show high robustness against noise, in real environments the recognition performance decreases markedly due to the effect of the Lombard reflex. In this paper, we also report experimental results showing the negative impact effect of the Lombard reflex on non-audible murmur recognition. In addition to a dictation task, we also report a keyword-spotting system based on non-audible murmur with very promising results.

1. Introduction

Non-Audible murmur (NAM) is very quietly uttered speech that cannot be heard by listeners near the talker. It is captured using a NAM microphone [1], which is a special acoustic sensor attached behind the talker's ear. Figure 1 shows the design of Silicon NAM microphone developed by Nakajima et al. in *Nara Institute of Science and Technology, Japan*. A NAM microphone is a body-conductive acoustic transducer, in which speech is captured directly from the talker's body through tissue or bone. Thus, such a transducer shows high robustness against noise and can capture voices with a very low intensity. Similar studies have been proposed by Zheng et al. [2], Graciarena et al. [3], and Jou et al [4] for noise robust speech recognition or soft whisper speech recognition.

Similarly to whisper speech, non-audible 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 acts as a low-pass filter and the high-frequency components are attenuated. However, the non-audible murmur spectral components still provide sufficient information to distinguish and recognize sounds accurately. To realize this, new hidden Markov models (HMMs) have to be trained using non-audible murmur data.

Previously, we reported HMM-based non-audible murmur automatic recognition using a Stethoscope NAM microphone with very promising results [5]. We also reported experiments for integrated non-audible murmur recognition and audible speech recognition using a NAM microphone [6].

In this paper, we also investigate non-audible murmur recognition in noisy environments using a Stethoscope and a Silicon NAM microphone. However, because of the nature of non-audible murmur (e.g., privacy), it is of high importance to also deal with noisy conditions, such as background speech and office noise, in automatic non-audible recognition. We carried out experiments using simulated noisy test data and data recorded under noisy conditions. Although using simulated noisy data the performance did not decrease significantly compared with that of the clean case, using real noisy data the performance decreased markedly. To investigate this problem, we studied the role of the Lombard reflex [7, 8] in non-audible murmur recognition and conducted experiments using Lombard non-audible murmur data. Results showed, that the Lombard reflex seriously affects the performance of non-audible murmur.

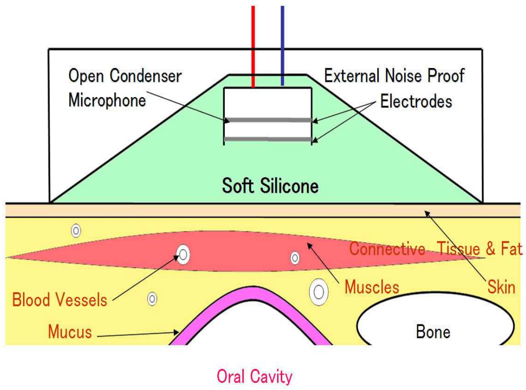

Figure 1: Silicon NAM microphone

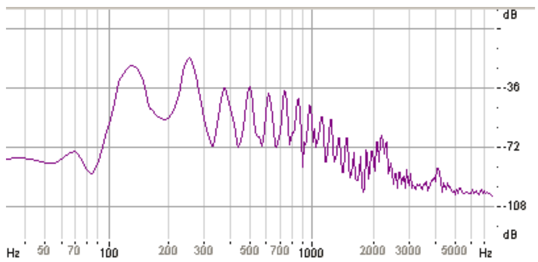

Figure 2: Power spectrum of the Japanese syllables /kini/ captured by NAM microphone

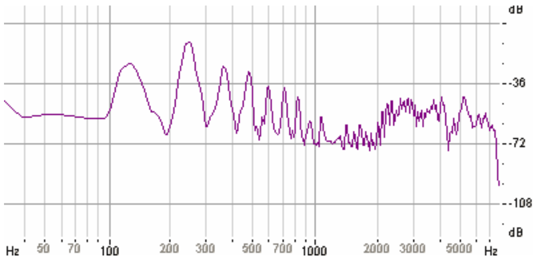

Figure 3: Power spectrum of the Japanese syllables /kini/ captured by close-talking microphone

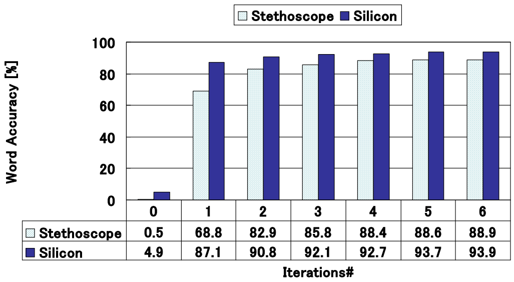

Figure 4: Non-audible murmur recognition using clean test data

2. Speaker-dependent non-audible murmur recognition

In this section, we present experimental results for speaker-dependent non-audible murmur recognition using NAM microphones. The recognition engine was the Julius 20k vocabulary Japanese dictation Toolkit [9]. The initial models were speaker-independent, gender-independent, 3000-state Phonetic Tied Mixture (PTM) HMMs, trained with the JNAS database and the feature vectors were of length 25 (12 MFCC, 12 ΔMFCC, ΔE). The non-audible murmur HMMs were trained using a combination of supervised 128-classes regression tree MLLR [10] and MAP [11] adaptation methods. Using, however, MLLR and MAP combination, the parameters are previously transformed using MLLR, and the transformed parameteres 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 previously during MLLR. Due to the large difference between the training data and the initial models, single-iteration adaptation is not effective in non-audible murmur recognition. Instead, a multi-iteration adaptation scheme was used. The initial models are adapted using the training data and 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 showed, that after 5-6 iterations significant improvement was achieved compared to the single-iteration adaptation. This training procedure is similar to that proposed by [12], but the object is different.

2.1 Non-audible murmur recognition using clean 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 non-audible murmur utterances 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 a word performance comparable to normal-speech recognition (96.2% word accuracy). 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 non-audible 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.

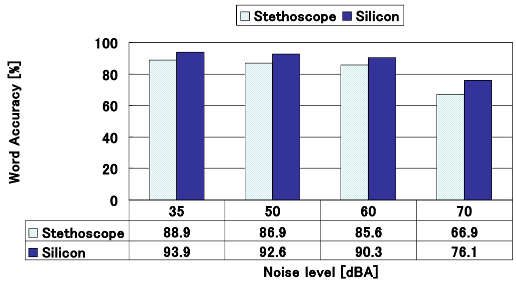

Figure 5: Non-audible murmur recognition using simulated noisy test data

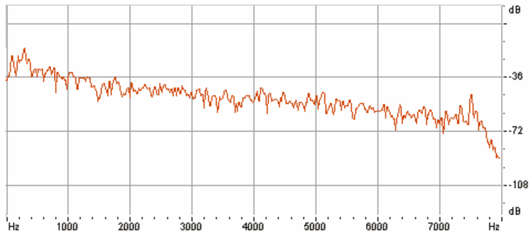

Figure 6: Long-term power spectrum of office noise used in our experiments

2.2 Non-audible murmur recognition using simulated noisy test data

In this experiment, office noise was played back at different levels (dBA) and recorded using a NAM microphone attached to a female talker. We recorded noises at 50 dBA and 60 dBA levels. The recorded noises were then superimposed on 24 clean non-audible murmur utterances, uttered by the same female speaker, to create the simulated noisy data. The acoustic models were trained using 100 non-audible murmur utterances recorded in a clean environment.

The results showed that the performance remained almost equal to that of the clean case when noise was superimposed on clean test data and recognition was performed using clean HMMs. More specifically, we achieved 83.7%, 82.9% and 80.9% word accuracies for the clean case, the 50 dBA noise

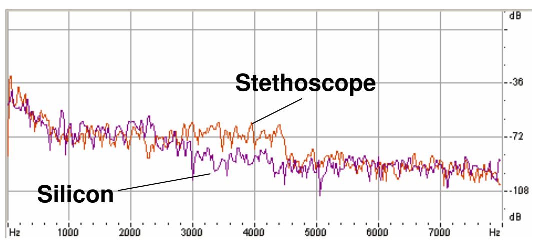

Figure 7: Long-term power spectrum of office noise at 70dBA level captured by NAM microphones

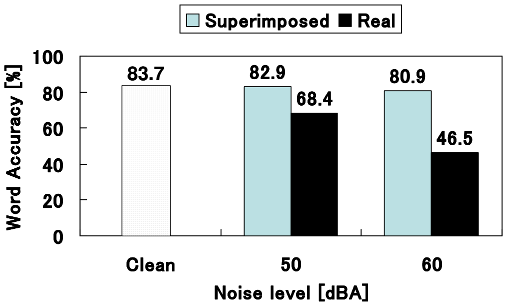

Figure 8: Non-Audible murmur recognition using noisy test data (office noise)

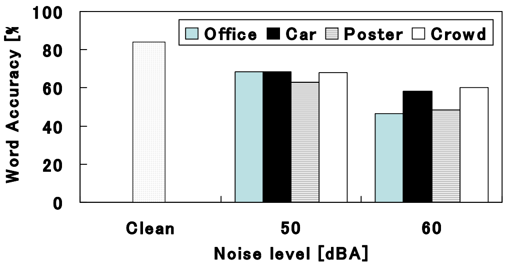

Figure 9: Non-Audible murmur recognition using various types of noise

level, and the 60 dBA noise level, respectively.

2.3 Non-audible murmur recognition using real noisy test data

In this section, we report experimental results for non-audible murmur recognition using real noisy database. The noisy test data were recorded in an environment, where different types of noise were playing back at 50 dBA and 60 dBA levels, while a speaker was uttering the test data. Four types of noise were used (office, car, poster, and crowd). For each noise and each level 24 utterances were recorded.

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 47%.

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 and office noises, the difference is larger.

Although, the performance using real noisy data is not markedly low and non-audible recognition is still possible, further investigations are necessary. In several studies, a negative impact effect of the Lombard reflex on automatic recognizers for normal speech has been reported. It is possible, therefore, that the degradations in word accuracy for non-audible murmur recognition when using real noisy data, are also related to the Lombard reflex. To realize this, we also addressed the Lombard reflex problem.

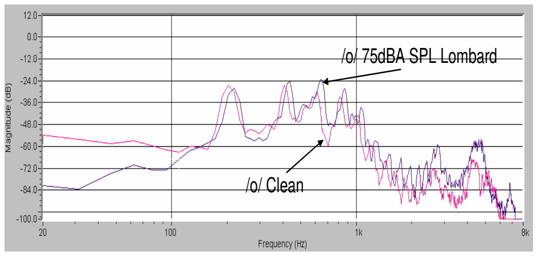

Figure 10: Power spectrum of clean vowel /O/ and Lombard vowel /O/

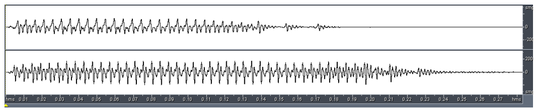

Figure 11: Waveform of clean vowel /O/ (upper) and Lombard vowel /O/

3. The role of the Lombard reflex in non-audible 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 while the speaker listens to noise through headphones or earphones. Even, 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 non-audible utterance recorded at 80 dBA. As can be seen, this Lombard non-audible murmur speech has characteristics similar to those of normal speech. Therefore, when non-audible murmur recognition is performed in noisy environments, the produced non-audible murmur characteristics are different than those of the non-audible murmur used in the training. As a result, the performance is degraded, even though the NAM microphone can capture non-audible murmur without a high sensitivity to environmental noise.

To show the effect of the Lombard reflex on non-audible murmur recognition, we carried out an experiment using Lombard non-audible murmur test data. The data were recorded in an anechoic room, while the speaker was listening to office noise through headphones. Since we used high-quality head-

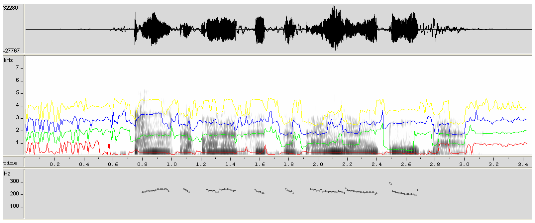

Figure 12: Lombard non-audible murmur recorded at 80 dBA

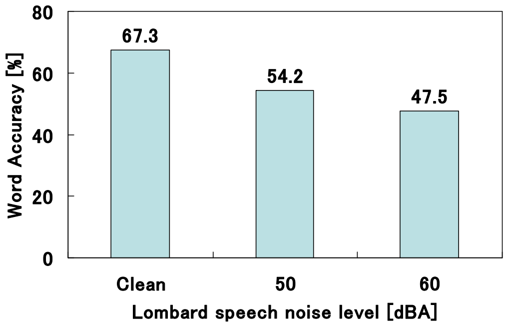

Figure 13: Non-Audible murmur recognition using Lombard data

phones, 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 non-audible murmur data using 50 utterances and MLLR adaptation.

Figure 13 shows the obtained results and the effect of the Lombard reflex on non-audible 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 non-audible murmur phenomena, the Lombard reflex is also present when there is no masking noise. However, due to the very low intensity of non-audible murmur, speakers might not hear their own voice. To make their voice audible, they increase their vocal levels, and as a result, non-audible murmur

4. A keyword-spotting system based on non-audible murmur

In this section, we present a keyword-spotting experiment for non-audible murmur. A non-audible murmur-based keyword-spotting system, however, can be applied to extract a specific number of keywords from unconstrained input speech in pri-

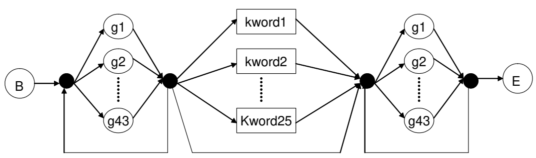

Figure 14: Grammar used in the keyword-spotter

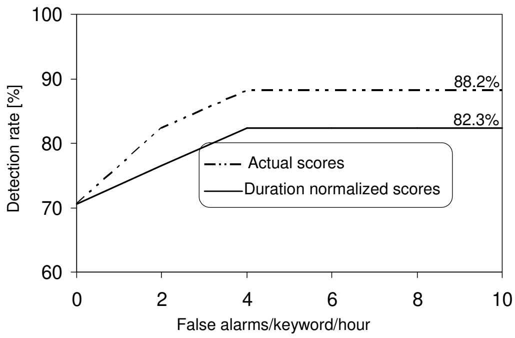

Figure 15: Receiver Operating Characteristics (ROC) for non-audible murmur keyword-spotter

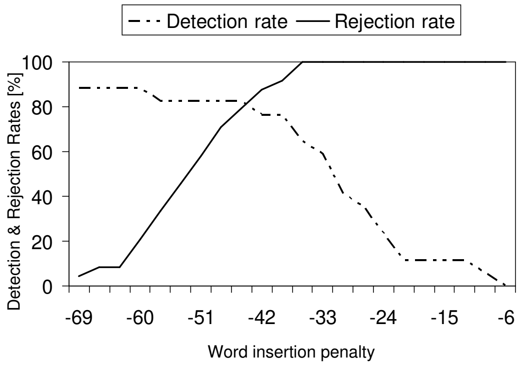

Figure 16: Detection and rejection rates for non-audible murmur keyword-spotter

vacy conditions. In some applications, when only a small number of keywords is required, a keyword-spotting system, with lower complexity and faster decoding, might be more reasonable than a dictation system.

In a keyword-spotting approach, not only the keywords, but also the non-keyword intervals must be modeled explicitly. Our approach, was based on phonemic garbage models [13]. The keywords were modeled using context-dependent HMMs, and monophone HMMs were used to model the non-keyword portions. Both HMM sets were trained with non-audible murmur data recorded using a silicon NAM microphone. Fourty-three monophone HMMs were connected as to allow any sequence. The vocabulary consisted of 25 keywords randomly selected from JNAS database. Figure 14 shows the grammar used in our experiment, which allowed at most one keyword per utterance.

In our experiment, the following evaluation measures were used:

- *Detection rate*. The percentage of keywords detected.
- *Rejection rate*. The percentage of non-keywords rejected.
- *Receiver Operating Characteristics (ROC) and Figure of Merit (FOM)*. The putative hits are sorted with respect to their scores, and the probability of detection at each false alarm is computed. The FOM is calculated as the average probability of detection between 0 and 10 false alarms per keyword.

For testing, we used 18 utterances, which included one keyword, and 24 utterances which did not include any keyword. Figure 15 shows the ROC curves. The figure shows, that by allowing 4 alarms per keyword we achieved 88.2% detection rate. The achieved FOM was 85.6%, which is promising result. The figure also shows, that using duration normalized scores the performance was decreased. Figure 16 shows the detection and rejection rates. To achieve higher detection and rejection rates, a word insertion penalty is tuned to decrease the likelihood of the garbage models. Without this tuning, however, a large number of false rejections (keywords are hypothesized as garbage models) appears, and as a result the detection rate decreases. With word insertion penalty tuning, we achieved a 82.5% equal rate (equal detection and rejection rates).

5. Conclusions

In this paper, we presented non-audible murmur recognition in clean and noisy 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. Non-Audible murmur recognition can be used when privacy in human-machine communication is desired. Since non-audible 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 non-audible 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 non-audible murmur recognition and we carried out an experiment using Lombard data. The results showed that the Lombard reflex has a negative impact effect on non-audible murmur recognition. Due to the speech production modifications, the non-audible murmur characteristics under Lombard conditions are changed and show a high similarity to normal speech. Due to this fact, a mismatch appears between the training and testing conditions and the performance decreases. As future work, we plan to investigate methods of decreasing the effect of the Lombard reflex on non-audible murmur recognition. A possible solution might be the adaptation of clean acoustic models to several Lombard conditions. In addition to a dictation task, we also reported a keyword-spotting experiment based on non-audible murmur with very promising results.

6. 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] Y. Zheng, Z. Liu, Z. Shang, M. Sinclair, J. Droppo, L. Deng, A. Acero, Z. Huang, " Air- and Bone-Conductive Integrated Microphones for Robust Speech Detection and Enhancement", *Proceedings of ASRU*, pp. 249–253, 2003.

[3] M. Graciarena, H. Franco, K. Sonmez, H. Bratt, "Combining Standard and Throat Microphones for Robust Speech Recognition", *IEEE Signal Processing Letters*, Vol. 10, No 3, pp.72–74, 2003.

[4] S. C. Jou, T. Schultz, Alex Weibel, "Adaptation for Soft Whisper Recognition Using a Throat Microphone", *Proceedings of ICSLP*, pp. –, 2004.

[5] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, K. Shikano, " Non-Audible Murmur (NAM) Recognition Using a Stethoscopic NAM microphone", *Proceedings of ICLP*, pp. 1469–1472, 2004.

[6] P. Heracleous, Y.Nakajima, A. Lee, H. Saruwatari, K. Shikano, "Audible (normal) speech and inaudible murmur recognition using NAM microphone", *Proceedings of EUSIPCO*, pp. 329–332, 2004.

[7] Junqua J-C, "The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizers", *J. Acoust. Soc. Am.*, Vol. 1 pp. 510–524, 1993.

[8] A. Wakao, K. Takeda, F. Itakura, "Variability of Lombard Effects Under Different Noise Conditions", *Proceedings of ICSLP*, pp. 2009–2012, 1996.

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

[10] 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.

[11] 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 Signal Processing*, Vol. 39, pp. 806–814, 1991.

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

[13] R. C. Rose, D. B. Paul, "A Hidden Markov Model Based Keyword Recognition System," *Proc. ICASSP*, pages 129–132, 1990.