ACOUSTIC MODEL TRAINING FOR NON-AUDIBLE MURMUR RECOGNITION USING TRANSFORMED NORMAL SPEECH DATA

Denis Babani, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano

Nara Institute of Science and Technology, Takayama 8916-5 Ikoma city, Nara Prefecture, Japan

ABSTRACT

In this paper we present a novel approach to acoustic model training for non-audible murmur (NAM) recognition using normal speech data transformed into NAM data. NAM is extremely soft murmur, that is so quiet that people around the speaker can hardly hear it. It is detected directly through the soft tissue of the head using a special body-conductive microphone, NAM microphone, placed on the neck below the ear. NAM recognition is one of the promising silent speech interfaces for man-machine speech communication. We have previously shown the effectiveness of speaker adaptive training (SAT) based on constrained maximum likelihood linear regression (CMLLR) in NAM acoustic model training. However, since the amount of available NAM data is still small, the effect of SAT is limited. In this paper we propose modified SAT methods capable of using a larger amount of normal speech data by transforming them into NAM data. The experimental results demonstrate that the proposed methods yield an absolute increase of approximately 2% in word accuracy compared with the conventional method.

Index Terms— silent speech interfaces, non-audible murmur recognition, acoustic model, speaker adaptive training, transformed normal speech

1. INTRODUCTION

Nowadays the accuracy of speech recognition systems is sufficiently high to be used in daily tasks. Even though there is confidence in the reliability of these systems, it is still difficult to imagine people making use of these functionalities in everyday life. A feeling of discomfort or even embarrassment in talking to machines (such as phones and car), disrupting silence in quite places, and a lack of privacy are likely reasons why people may try to avoid such convenient and hands-free input interfaces.

Silent speech interfaces [1] have recently been studied as a technology to enable speech communication to take place without the necessity of emitting an audible acoustic signal. Various sensing devices, such as a throat microphone [2], electromyography (EMG) [3], and ultrasound imaging [4], have been explored as alternatives to air microphones. These sensing devices are effective for soft speech in private conversation and as a speaking aid for people with a vocal disability.

As a sensing device for silent speech, Nakajima *et al.* [5] developed a non-audible murmur (NAM) microphone, which is a special body-conductive microphone. Inspired by a stethoscope, the NAM microphone was originally developed to detect extremely soft murmur called NAM, which is so faint that people around the speaker can hardly hear it. Placed on the neck below the ear, a NAM microphone is capable of detecting various types of speech such as NAM, whisper, and normal speech through the soft tissue of the head. Moreover, it has greater usability than other devices such as EMG and ultrasound systems.

NAM recognition systems are not very different from those utilizing normal speech. In fact, language models, dictionaries, searching algorithms, and other specific modules may be used without any modifications at all. The only modifications required are in the acoustic model, which should match the acoustic features of NAM. However if we built a normal speech acoustic model for NAM, it would take many years to gather sufficient training data, and obtain satisfactory accuracy in NAM recognition. One possible shortcut is to use currently existing normal speech databases. As reported in [6, 7], normal speech data can be used to generate an initial acoustic model, then model adaptation techniques (e.g., [8]) can be applied to it to develop a speaker-dependent NAM acoustic model using a small amount of NAM data. It was been also reported in [9] that speaker adaptive training (SAT) [10] yields significant improvements in NAM recognition accuracy by refining the initial acoustic model using only the NAM data of several tens of speakers.

In this paper we propose a novel approach to NAM acoustic model training to further increase the accuracy of the NAM acoustic model. Some of the canonical model parameters updated in the conventional SAT are not well optimized since the available NAM data are still limited. Inspired by a speech synthesis technique for transforming NAM into normal speech [11], the proposed method transforms acoustic features of normal speech into those of NAM to effectively increase the amount of NAM data available in SAT. This is achieved by modifying the SAT process on the basis of constrained maximum likelihood linear regression (CMLLR) [8]. The experimental results of the proposed methods indicate an increase in absolute word accuracy of approximately 2% compared with the conventional method.

This paper is organized as follows. In section 2 we give a short description of NAM. In section 3, previous work on NAM recognition including SAT for NAM and limitations of this approach are described. In section 4 we explain the proposed method in more detail, which is followed by its evaluation in section 5. Finally, we summarize this paper in section 6.

2. NON-AUDIBLE MURMUR (NAM)

NAM is defined as the articulated production of respiratory sound without using the vibration of vocal folds. It is modulated by various acoustic filter characteristics as a result of the motion and interaction of speech organs, and is transmitted through the soft tissues of the human body [5]. NAM can be detected with a NAM microphone attached on the surface of the human body. According to Nakajima *et al.*, the optimal position for a NAM microphone is behind the ear.

The sampled signal is weak and is amplified before analysis by speech recognition tools. The amplified NAM is still fairly intelligible and its sound quality is unnatural since high frequency components over 3 or 4 kHz are severely attenuated by the features of body conduction such as the lack of radiation from the lips and the effect of the low-pass characteristics of the soft tissue.

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

3. DEVELOPMENT OF NAM ACOUSTIC MODEL

3.1. Previous Work

NAM utterances recorded with a NAM microphone can be used to train speaker-dependent hidden Markov models (HMMs) for NAM recognition. The simplest way to build a NAM acoustic model would be to start from scratch and utilize only NAM samples. However, this method would require a large amount of training data, which is not available for NAM.

Another method of building a NAM acoustic model would be to retrain a speaker-independent normal speech model using NAM samples. This method requires less training data compared with training from scratch. In [6] it was reported that an iterative MLLR adaptation process using the adapted model as the initial model in the next EM (expectation-maximization algorithm) iteration step is very effective because the acoustic characteristics of NAM are considerably different from those of normal speech.

We previously demonstrated that the use of a canonical model for NAM adaptation that is trained using NAM data in the SAT paradigm yields significant improvements in the performance of NAM recognition [9]. A schematic representation of this method is shown in figure 1. In CMLLR-based SAT, the speaker-dependent CMLLR transform $\boldsymbol{W}_n^{(NAM)} = \left[\boldsymbol{b}_n^{(NAM)}, \boldsymbol{A}_n^{(NAM)}\right]$ is applied to the feature vector $\boldsymbol{o}_t^{(n)}$ as follows:

$$\hat{\boldsymbol{o}}_t^{(n)} = \boldsymbol{A}_n^{(NAM)}\boldsymbol{o}_t^{(n)} + \boldsymbol{b}_n^{(NAM)} = \boldsymbol{W}_n^{(NAM)}\boldsymbol{\zeta}_t^{(n)}, \quad (1)$$

where $n \in \{1, \cdots, N\}$ and $t \in \{1, \cdots, T_n\}$ are indexes for the NAM speaker and time, respectively, and $\boldsymbol{\zeta}_t^{(n)}$ is the extended feature vector $\left[1, \boldsymbol{o}_t^{(n)\top}\right]^\top$. The auxiliary function of the EM algorithm in SAT is given by

$$\mathcal{Q}\left(\left\{\boldsymbol{\lambda}, \boldsymbol{W}_{1:N}^{(NAM)}\right\}, \left\{\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{W}}_{1:N}^{(NAM)}\right\}\right) \propto -\frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\sum_{m=1}^{M}\gamma_{m,t}^{(n)}\mathcal{L}_{n,m,t}^{(NAM)}, \quad (2)$$

where $m \in \{1, \cdots, M\}$ is an index of Gaussian component, $\boldsymbol{W}_{1:N}^{(NAM)}$ is the set of speaker-dependent CMLLR transforms $\left\{\boldsymbol{W}_1^{(NAM)}, \cdots, \boldsymbol{W}_N^{(NAM)}\right\}$, and

$$\mathcal{L}_{n,m,t}^{(NAM)} = \log|\hat{\boldsymbol{\Sigma}}_m| - \log\left|\hat{\boldsymbol{A}}_n^{(NAM)}\right|^2 + \left(\hat{\boldsymbol{W}}_n^{(NAM)}\boldsymbol{\zeta}_t^{(n)} - \hat{\boldsymbol{\mu}}_m\right)^\top \hat{\boldsymbol{\Sigma}}_m^{-1}\left(\hat{\boldsymbol{W}}_n^{(NAM)}\boldsymbol{\zeta}_t^{(n)} - \hat{\boldsymbol{\mu}}_m\right). \quad (3)$$

In the E-step, $\gamma_{m,t}^{(n)}$ is calculated as the posterior probability of component m generating feature vector $\boldsymbol{o}_t^{(n)}$ given the current model parameter set $\boldsymbol{\lambda}$, the CMLLR transform set $\boldsymbol{W}_{1:N}^{(NAM)}$, and the feature vector sequence $\left\{\boldsymbol{o}_1^{(n)}, \cdots, \boldsymbol{o}_{T_n}^{(n)}\right\}$. In the M-step, the updated model parameter set $\hat{\boldsymbol{\lambda}}$ including the mean vector $\hat{\boldsymbol{\mu}}_m$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_m$ of each Gaussian component and the updated CMLLR transform set $\hat{\boldsymbol{W}}_{1:N}^{(NAM)}$ are sequentially determined by maximizing the auxiliary function. The initial model parameter set for SAT is set to that of a speaker-independent model developed using normal speech data sets consisting of voices of several hundred speakers. Finally, a speaker-dependent model for individual speakers is developed from the canonical model using iterative MLLR mean and variance adaptation.

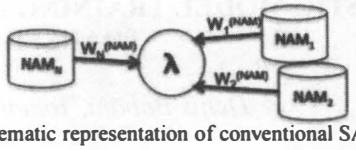

Fig. 1. Schematic representation of conventional SAT process.

Note that multiple linear transforms are used for each speaker. The Gaussian components are automatically clustered according to the amount of adaptation data using a regression-tree-based approach [12].

3.2. Problem

Even though the conventional SAT method produces some improvement in recognition accuracy, further improvements are essential for the development of a NAM recognition interface. One of the problems in this method continues to be the limitation of training data. This is a serious problem when using a normal speech acoustic model including many HMM model parameters as the starting point. Although such a complicated acoustic model is well adapted to NAM data in MLLR or CMLLR adaptation since all Gaussian components are transformed by effectively sharing the same linear transform among different components, it generates one issue in the development of the canonical model. Since each Gaussian component is updated using component-dependent sufficient statistics calculated from NAM data, there are many components that are not well updated due to the lack of training data. Consequently, the effectiveness of SAT is reduced or lost for such components, adversely affecting the adaptation performance.

4. IMPROVING NAM ACOUSTIC MODEL USING TRANSFORMED NORMAL SPEECH DATA

4.1. Proposed SAT Using Transformed Normal Speech Data

A schematic representation of the proposed method is shown in figure 2. To normalize acoustic variations caused by both speaker differences and speaking style differences (i.e., differences between NAM and normal speech), the speaker-dependent CMLLR transform $\boldsymbol{W}_s^{(S2N)} = \left[\boldsymbol{b}_s^{(S2N)}, \boldsymbol{A}_s^{(S2N)}\right]$ is applied to the feature vector $\boldsymbol{o}_t^{(s)}$ of normal speech as follows:

$$\hat{\boldsymbol{o}}_t^{(s)} = \boldsymbol{A}_s^{(S2N)}\boldsymbol{o}_t^{(s)} + \boldsymbol{b}_s^{(S2N)} = \boldsymbol{W}_s^{(S2N)}\boldsymbol{\zeta}_t^{(s)}, \quad (4)$$

where $s \in \{1, \cdots, S\}$ is the index for a speaker of normal speech. The auxiliary function in the proposed method is given by

$$\mathcal{Q}\left(\left\{\boldsymbol{\lambda}, \boldsymbol{W}_{1:N}^{(NAM)}, \boldsymbol{W}_{1:S}^{(S2N)}\right\}, \left\{\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{W}}_{1:N}^{(NAM)}, \hat{\boldsymbol{W}}_{1:S}^{(S2N)}\right\}\right) \propto -\frac{1}{2}\left(\sum_{n=1}^{N}\sum_{t=1}^{T_n}\sum_{m=1}^{M}\gamma_{m,t}^{(n)}\mathcal{L}_{n,m,t}^{(NAM)} + \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{m=1}^{M}\gamma_{m,t}^{(s)}\mathcal{L}_{s,m,t}^{(SP)}\right), \quad (5)$$

where $\boldsymbol{W}_{1:S}^{(S2N)}$ is the set of speaker-dependent CMLLR transforms for normal speech $\left\{\boldsymbol{W}_1^{(S2N)}, \cdots, \boldsymbol{W}_S^{(S2N)}\right\}$, and

$$\mathcal{L}_{s,m,t}^{(SP)} = \log|\hat{\boldsymbol{\Sigma}}_m| - \log\left|\hat{\boldsymbol{A}}_s^{(S2N)}\right|^2 + \left(\hat{\boldsymbol{W}}_s^{(S2N)}\boldsymbol{\zeta}_t^{(s)} - \hat{\boldsymbol{\mu}}_m\right)^\top \hat{\boldsymbol{\Sigma}}_m^{-1}\left(\hat{\boldsymbol{W}}_s^{(S2N)}\boldsymbol{\zeta}_t^{(s)} - \hat{\boldsymbol{\mu}}_m\right). \quad (6)$$

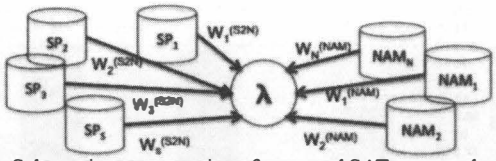

Fig. 2. Schematic representation of proposed SAT process described in section 4.1.

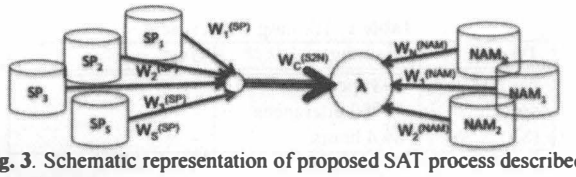

Fig. 3. Schematic representation of proposed SAT process described in section 4.2.

In the E-step, the posterior probabilities $\gamma_{m,t}^{(n)}$ and $\gamma_{m,t}^{(s)}$ are calculated from the current model parameter set $\boldsymbol{\lambda}$ and the CMLLR transform sets $\boldsymbol{W}_{1:N}^{(NAM)}$ and $\boldsymbol{W}_{1:S}^{(S2N)}$. In the M-step, the model parameter set and the CMLLR transform sets are sequentially updated. The initial model parameter set for SAT is set to that of the canonical model developed by the conventional SAT process described in section 3.1. Multiple linear transforms are used for each speaker.

4.2. Proposed SAT with Factorized Transforms

Because the acoustic characteristics of NAM are considerably different from those of normal speech, a more complicated transformation will be effective for transforming the normal speech data of different speakers into the NAM data of a canonical speaker. Such a complicated transformation can be achieved by increasing the number of linear transforms, but the estimation accuracy of the linear transforms will suffer from a decrease in the amount of adaptation data available for the estimation of each transform. To make it possible to effectively increase the number of linear transforms while maintaining a sufficiently high estimation accuracy, factorized transforms are applied in the proposed method.

A schematic representation of the proposed method using the factorized transforms is shown in figure 3. The CMLLR transform $\boldsymbol{W}_s^{(S2N)} = \left[\boldsymbol{b}_s^{(S2N)}, \boldsymbol{A}_s^{(S2N)}\right]$ is factorized into two CMLLR transforms: one is a speaker-dependent transform in normal speech, $\boldsymbol{W}_s^{(SP)} = \left[\boldsymbol{b}_s^{(SP)}, \boldsymbol{A}_s^{(SP)}\right]$, and the other is a speaker-independent style transform from normal speech to NAM, $\boldsymbol{W}_c^{(S2N)} = \left[\boldsymbol{b}_c^{(S2N)}, \boldsymbol{A}_c^{(S2N)}\right]$. The factorized transforms are applied to the feature vector of normal speech as follows:

$$\hat{\boldsymbol{o}}_t^{(s)} = \boldsymbol{A}_c^{(S2N)}\left(\boldsymbol{A}_s^{(SP)}\boldsymbol{o}_t^{(s)} + \boldsymbol{b}_s^{(SP)}\right) + \boldsymbol{b}_c^{(S2N)} = \boldsymbol{W}_{c,s}^{(S2N)}\boldsymbol{\zeta}_t^{(s)}, \quad (7)$$

where the composite transform $\boldsymbol{W}_{c,s}^{(S2N)}$ is represented as $\left[\boldsymbol{A}_c^{(S2N)}\boldsymbol{b}_s^{(SP)} + \boldsymbol{b}_c^{(S2N)}, \boldsymbol{A}_c^{(S2N)}\boldsymbol{A}_s^{(SP)}\right]$. The auxiliary function in the proposed method using the factorized transforms is given by

$$\mathcal{Q}\left(\left\{\boldsymbol{\lambda}, \boldsymbol{W}_{1:N}^{(NAM)}, \boldsymbol{W}_{1:S}^{(SP)}, \boldsymbol{W}_c^{(S2N)}\right\}, \left\{\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{W}}_{1:N}^{(NAM)}, \hat{\boldsymbol{W}}_{1:S}^{(SP)}, \hat{\boldsymbol{W}}_c^{(S2N)}\right\}\right)$$
$$\propto -\frac{1}{2}\left(\sum_{n=1}^{N}\sum_{t=1}^{T_n}\sum_{m=1}^{M}\gamma_{m,t}^{(n)}\mathcal{L}_{n,m,t}^{(NAM)} + \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{m=1}^{M}\gamma_{m,t}^{(s)}\mathcal{L}_{s,m,t}^{\prime(SP)}\right), \quad (8)$$

where

$$\mathcal{L}_{s,m,t}^{\prime(SP)} = \log|\hat{\boldsymbol{\Sigma}}_m| - \log\left|\hat{\boldsymbol{A}}_s^{(SP)}\right|^2 - \log\left|\hat{\boldsymbol{A}}_c^{(S2N)}\right|^2 + \left(\hat{\boldsymbol{W}}_{c,s}^{(S2N)}\boldsymbol{\zeta}_t^{(s)} - \hat{\boldsymbol{\mu}}_m\right)^{\top}\hat{\boldsymbol{\Sigma}}_m^{-1}\left(\hat{\boldsymbol{W}}_{c,s}^{(S2N)}\boldsymbol{\zeta}_t^{(s)} - \hat{\boldsymbol{\mu}}_m\right). \quad (9)$$

Multiple linear transforms are used for each speaker and for the speaker-independent style transformation. The canonical model developed by the conventional SAT process described in section 3.1 is used as the initial model. The speaker-dependent transforms in normal speech, $\boldsymbol{W}_s^{(SP)}$, are initialized by the conventional SAT process using only normal speech data, where the speaker-independent normal speech model is used as the initial model. Int his paper, $\boldsymbol{W}_s^{(SP)}$ are fixed to the initialized parameters throughout the proposed SAT process. They may also be updated iteratively.

Note that the number of style transforms is easily increased since all normal speech data are effectively used for their estimation. Consequently, a larger number of composite transforms is available, than the number of speaker-dependent transforms available in the other proposed SAT process described in section 4.1.

4.3. Implementation

We have found that if both normal speech data and NAM data are used simultaneously to update the canonical model parameters, the NAM recognition accuracy of the speaker-dependent adaptation model generated from the updated canonical model tends to decrease considerably. This is because the proposed method does not perfectly map normal speech features into NAM features and the canonical model matches normal speech features better than NAM features due to the use of a much larger amount of normal speech data than NAM data.

To avoid this issue, in this paper the transformed normal speech data are only used to develop the first canonical model, then, this model is further updated in SAT using only NAM data. Namely, after optimizing the speaker-dependent linear transform set $\boldsymbol{W}_{1:S}^{(S2N)}$ or the style transforms $\boldsymbol{W}_c^{(S2N)}$ while fixing the model parameters to the initial values (i.e., the canonical model parameters optimized in conventional SAT using NAM data), the model parameters are updated using only transformed normal speech data by maximizing the part of the auxiliary function related to $\mathcal{L}_{s,m,t}^{(SP)}$ in Eq. (5) or $\mathcal{L}_{s,m,t}^{\prime(SP)}$ in Eq. (8). The model parameters are finally updated in the SAT process using only NAM data by maximizing the part of the auxiliary function related to $\mathcal{L}_{n,m,t}^{(NAM)}$. In this implementation, the proposed methods are only different from the conventional method in that the initial model parameters in SAT with NAM are developed using the transformed normal speech data.

5. EXPERIMENTAL EVALUATION

5.1. Experimental Conditions

Table 1 shows the amount training and test data. The starting acoustic model was a speaker-independent (SI) three-state left-to-right tied-state triphone HMM for normal speech, for which each state output probability density was modeled by a Gaussian mixture model (GMM) with 16 mixture components. The total number of triphones was 3300. The employed acoustic feature vector was a 25-dimensional vector including 12 MFCC, 12 Δ MFCC, and Δ Energy. A dictionary of approximately 63 k words (multiple pronunciations) and a bigram language model were used during decoding.

Table 1. Training and test sets

Type	Training	Test
Normal speech (SP)	298 speakers 46980 utterances 84.4 hours	- - -
NAM	42 speakers 8893 utterances 15.5 hours	41 speakers 1023 utterances 1.83 hours

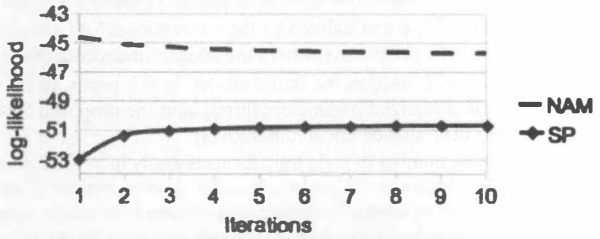

Fig. 4. Change in log-scaled likelihoods for training utterances.

The regression-tree based approach was adopted to dynamically determine the regression classes used to estimate multiple CMLLR transforms. In the SAT process, the average numbers of speaker-specific linear transforms for normal speech and for NAM were approximately 104 and 110, respectively. The number of style transforms from normal speech to NAM was manually set to 256.

5.2. Experimental Results

To illustrate the implementation issue described in section 4.3, the proposed SAT with the factorized transforms was performed using both NAM data and normal speech data to update the canonical model. Figure 4 shows the change in log-likelihoods of the training utterances of NAM and normal speech with the number of adaptive iterations in the SAT process. In each iteration the NAM speaker-dependent CMLLR transforms and style transforms were calculated, and then the canonical model was updated. It can be observed from this figure that during the iterative estimation, the likelihood for normal speech data tends to increase while that for NAM data tends to decrease. Consequently, the resulting canonical model caused the degradation of NAM recognition accuracy.

To demonstrate the effectiveness of the proposed methods, the canonical models were developed by the proposed SAT methods based on the implementation in section 4.2 and the conventional SAT method, and then the speaker-dependent models were built from each canonical model using the CMLLR adaptation. Figure 5 shows the results with a 5% confidence level. The proposed methods yield significant improvements in word accuracy (WACC) compared with the conventional method. We found that 1115 triphone models (approximately 1/3 of the HMM set) were not observed in the NAM training data. The canonical model parameters in these states were not updated at all in the conventional SAT. On the other hand, they were updated in the proposed methods using the transformed normal speech data. This is one of the major factors yielding the improvement in WACC shown in figure 5. Moreover, it can also be observed that the use of the factorized transformations yields a slight improvement in the proposed method.

[1] These experimental conditions are different from those in [9].

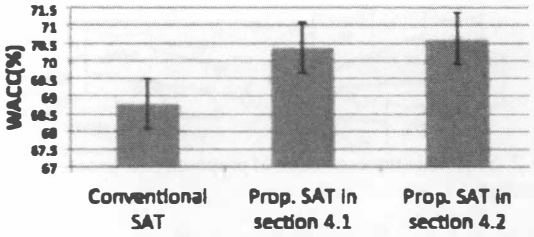

Fig. 5. Word accuracy of different methods.

6. CONCLUSIONS

In this paper, we proposed modified speaker adaptive training (SAT) methods for building a canonical model for non-audible murmur (NAM) adaptation so as to make available a larger amount of normal speech data transformed into NAM data in the training. The experimental results demonstrated that the proposed methods yield significant improvement in NAM recognition accuracy compared with the conventional SAT method since it is capable of extracting more information from normal speech data and applying it to the training process of the NAM acoustic model. Moreover, the use of factorized transformations in the proposed method yields a slight improvement in the performance of NAM recognition. A further investigation will be conducted on regression tree generation in the SAT process.

7. REFERENCES

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, and J.S. Brumberg. Silent speech interfaces. *Speech Communication*, Vol. 52, No. 4, pp. 270–287, 2010.

[2] S-C. Jou, T. Schultz, and A. Waibel. Adaptation for soft whisper recognition using a throat microphone. *Proc. INTERSPEECH*, pp. 1493–1496, Jeju Island, Korea, 2004.

[3] T. Schultz and M. Wand. Modeling coarticulation in EMG-based continuous speech recognition. *Speech Communication*, Vol. 52, No. 4, pp. 341–353, 2010.

[4] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication*, Vol. 52, No. 4, pp. 288–300, 2010.

[5] Y. Nakajima, H. Kashioka, N. Cambell, and K. Shikano. Non-Audible Murmur (NAM) Recognition. *IEICE Trans. Information and Systems*, Vol. E89-D, No. 1, pp. 1–8, 2006.

[6] 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. *Proc. ASRU*, pp. 73–76, St. Thomas, USA, Dec. 2003.

[7] P. Heracleous, V.-A. Tran, T. Nagai, and K. Shikano. Analysis and recognition of NAM speech using HMM distances and visual information. *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 18, No. 6, pp. 1528–1538, 2010.

[8] M.J.F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, Vol. 12, No. 2, pp. 75–98, 1998.

[9] T. Toda, K. Nakamura, T. Nagai, T. Kaino, Y. Nakajima, and K. Shikano. Technologies for processing body-conducted speech detected with non-audible murmur microphone. *Proc. INTERSPEECH*, pp. 632–635, Brighton, UK, Sep. 2009.

[10] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. *Proc. ICSLP*, pp. 1137–1140, Philadelphia, Oct. 1996.

[11] T. Toda and K. Shikano. NAM-to-speech conversion with Gaussian mixture models. *Proc. INTERSPEECH*, pp. 1957–1960, Lisbon, Portugal, Sep. 2005.

[12] M.J.F. Gales. The generation and use of regression class trees for MLLR adaptation. *Technical Report*, CUED/F-INFENG/TR263, Cambridge University, 1996.