Vol.2011-SLP-85 No.2
2011/2/4

The Use of Transformed Normal Speech Data in Acoustic Model Training for Non-Audible Murmur Recognition

Denis Babani,[†1] Tomoki Toda,[†1]
Hiroshi Saruwatari[†1] and Kiyohiro Shikano[†1]

This paper presents a novel approach to the acoustic model training for Non-Audible Murmur (NAM) recognition using normal speech data transformed into NAM data. NAM is extremely soft murmur, which is so quiet that people around the speaker hardly hear it. NAM recognition is one of the promising silent speech interfaces for man-machine speech communication. Our previous work has shown the effectiveness of Speaker Adaptive Training (SAT) based on Constrained Maximum Likelihood Linear Regression (CMLLR) in the 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 transformation of normal speech data is performed with the CMLLR adaptation. The experimental results demonstrate that the proposed methods yield an absolute increase of around 2% in word accuracy compared with the conventional method.

1. Introduction

Nowadays accuracy of speech recognition systems is considerably high enough to be used in daily tasks. Even though there is such confidence in these systems, it is still hard to see people making use of these functionalities in everyday life. Feeling uncomfortable (even embarrassment) in talking to machines (phone, car, etc.), being noisy, and lack of privacy would be one of the main reasons why people try to avoid such convenient and hands-free input interfaces.

Silent speech interfaces[1)] have recently 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 alternative to air microphone. These sensing devices are effective for soft speech in a private talk and as a speaking aid for the vocally handicapped.

As one of the sensing devices for silent speech, Nakajima *et al.*[5)] have developed Non-Audible Murmur (NAM) microphone, which is a special body-conductive microphone. Inspired by a stethoscope, NAM microphone was originally developed to detect extremely soft murmur called NAM, which is so quiet that people around the speaker hardly hear it. Placed on the neck below the ear, 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, its usability is better compared with other devices such as EMG or ultrasound systems. Considering these properties, we focus on NAM microphone as one of the promising devices.

Building NAM recognition systems are not very different from those utilizing normal speech. As a matter of fact, language models, dictionaries, searching algorithms, and other specific modules may be used without any modifications at all. The only modification is needed in the acoustic model, which should match acoustic features of NAM. If we follow the same footsteps of building a normal speech acoustic model, it would take many years gathering training data to reach a considerable accuracy in NAM recognition. One possible shortcut is the use of already accumulated normal speech databases. As reported in the literatures[6),7)], normal speech data can serve for generating an initial acoustic model, and later model adaptation techniques (e.g., linear regression approaches[8)]) are applied to it for developing a speaker-dependent NAM acoustic model using a small amount of NAM data. It has been also reported inthe literature[9)] that speaker adaptive training (SAT)[10)] yields significant improvements in NAM recognition accuracy by refining the initial acoustic model using only several tens of speakers' NAM data.

In this paper we propose a novel approach to the NAM acoustic model training for further increasing the accuracy of NAM acoustic model. Some of canonical model parameters updated in the conventional SAT are not well optimized since the amount of available NAM data is still limited. Inspired by a speech synthesis

†1 Nara Institute of Science and Technology

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 proposed process is achieved by modifying the SAT process based on Constrained Maximum Likelihood Linear Regression (CMLLR)[8]. The experimental results of the proposed methods yield around 2% increase in absolute word accuracy compared to the conventional methods.

This paper is organized as follows. Section 2 gives a short description of NAM. In Section 3, conventional work on NAM recognition including SAT for NAM and the problem of this approach are described. Section 4 explains in more detail the proposed method, 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 vocal folds vibrations, modulated by various acoustic filter characteristics as a result of motion and interaction of speech organs, and transmitted through soft tissues of human body[5]. NAM can be detected with NAM microphone attached on the surface of human body. According to Nakajima *et. al.*, the optimal position for it would be just behind the ear.

The sampled signal is weak, and usually is amplified before analyzed by speech recognition tools. The amplified NAM is still less intelligible and its sound quality is unnatural since high frequency components over 3 or 4 kHz are severely attenuated by essential mechanisms of body conduction such as lack of radiation characteristics from lips and influence of low-pass characteristics of the soft tissue[12].

3. DEVELOPMENT OF NAM ACOUSTIC MODEL

3.1 Conventional Work

NAM utterances recorded with 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 requires a lot of training data, which is not the case for NAM.

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

We have previously demonstrated that the use of the canonical model for NAM adaptation trained using NAM data in SAT paradigm yields significant improvements in the performance of NAM recognition[9]. A schematic representation of this method is shown in figure 1. In the 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:

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

where $n \in \{1, \cdots, N\}$ and $t \in \{1, \cdots, T_n\}$ are indexes for 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

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

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

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

In E-step, $\gamma_{m,t}^{(n)}$ is calculated as the posterior probability of the component m generating the feature vector $\boldsymbol{o}_t^{(n)}$ given the current model parameter set $\boldsymbol{\lambda}$, the CMLLR transform $\boldsymbol{W}_n^{(NAM)}$, and the feature vector sequence $\left\{\boldsymbol{o}_1^{(n)}, \cdots, \boldsymbol{o}_{T_n}^{(n)}\right\}$ for

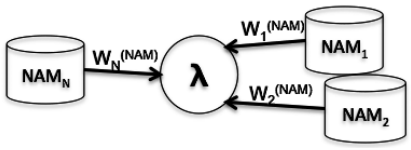

Fig. 1 Schematic representation of conventional SAT process.

each speaker. In M-step, the updated model parameter set $\hat{\boldsymbol{\lambda}}$ including a mean vector $\hat{\boldsymbol{\mu}}_m$ and a 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 with normal speech data sets consisting of voices of several hundreds of speakers. After the canonical model is optimized with SAT, the speaker-dependent model for individual speakers were developed from the canonical model using iterative MLLR mean and variance adaptation.

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[13].

3.2 Problem

Even though the conventional SAT method showed some improvements in the recognition accuracy, further improvements would be essential in the development of NAM recognition interface. One of the problems spotted in this method continues to be the limitation of training data. This is a serious problem when using a normal speech acoustic model as a starting point, which includes many HMM model parameters. The MLLR or CMLLR adaptation enables such a complicated acoustic model to be well adapted to NAM data since all Gaussian components are transformed by effectively sharing the same linear transform among different components. Therefore, the use of the complicated acoustic model is very effective in the adaptation. However, it causes one issue in the development of the canonical model. Since each Gaussian component of the canonical model is updated with component-dependent sufficient statistics calculated from NAM data, there are a lot of components not well updated due to lack of the amount of training data. Consequently, the effectiveness of SAT is minimized or lost in those components and the adaptation performance will suffer from them.

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:

$$\hat{\boldsymbol{o}}_t^{(s)} = \boldsymbol{A}_s^{(S2N)}\boldsymbol{o}_t^{(s)} + \boldsymbol{b}_s^{(S2N)} = \boldsymbol{W}_s^{(S2N)}\boldsymbol{\zeta}_t^{(s)}, \tag{3}$$

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

$$\begin{aligned}&\mathcal{Q}\left(\left\{\boldsymbol{\lambda}, \boldsymbol{W}_{1:N}^{(NAM)}, \boldsymbol{W}_{1:S}^{(S2N)}\right\}, \left\{\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{W}}_{1:N}^{(NAM)}, \hat{\boldsymbol{W}}_{1:S}^{(S2N)}\right\}\right)\\ &\propto -\frac{1}{2}\left(\sum_{n=1}^{N}\sum_{t=1}^{T_n}\sum_{m=1}^{M}\gamma_{m,t}^{(n)}\mathcal{L}_{n,m,t}^{(NAM)} + \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{m=1}^{M}\gamma_{m,t}^{(s)}\mathcal{L}_{s,m,t}^{(SP)}\right),\end{aligned} \tag{4}$$

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

$$\begin{aligned}\mathcal{L}_{s,m,t}^{(SP)} &= \log|\hat{\boldsymbol{\Sigma}}_m| - \log\left|\hat{\boldsymbol{A}}_s^{(S2N)}\right|^2\\ &+ \left(\hat{\boldsymbol{W}}_s^{(S2N)}\boldsymbol{\zeta}_t^{(s)} - \hat{\boldsymbol{\mu}}_m\right)^{\top}\hat{\boldsymbol{\Sigma}}_m^{-1}\left(\hat{\boldsymbol{W}}_s^{(S2N)}\boldsymbol{\zeta}_t^{(s)} - \hat{\boldsymbol{\mu}}_m\right).\end{aligned} \tag{5}$$

In E-step, the posterior probabilities, $\gamma_{m,t}^{(n)}$ and $\gamma_{m,t}^{(s)}$, for individual speakers are calculated given the current model parameter set $\boldsymbol{\lambda}$ and the CMLLR transform sets, $\boldsymbol{W}_{1:N}^{(NAM)}$ and $\boldsymbol{W}_{1:S}^{(S2N)}$. In 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

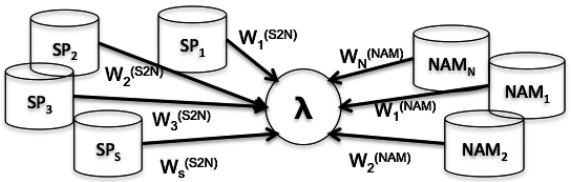

Fig. 2 Schematic representation of proposed SAT process described in section 4.1.

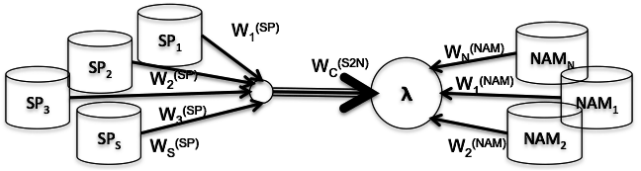

Fig. 3 Schematic representation of proposed SAT process described in section 4.2.

described in section 3.1. Multiple linear transforms are used for each speaker.

4.2 Proposed SAT with Factorized Transforms

Because acoustic characteristics of NAM are quite different from those of normal speech, more complicated transformation would be effective for transforming the normal speech data of different speakers into NAM data of a canonical speaker. Such a complicated transformation is achieved by an increase of the number of linear transforms but the estimation accuracy of linear transforms suffers from a decrease of 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 keeping the estimation accuracy high enough, factorized transforms are applied to the proposed method.

A schematic representation of the proposed method with 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 into 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:

$$\hat{\boldsymbol{o}}_t^{(s)} = \boldsymbol{A}_c^{(S2N)}\left(\boldsymbol{A}_s^{(SP)}\boldsymbol{o}_t^{(s)} + \boldsymbol{b}_s^{(SP)}\right) + \boldsymbol{b}_c^{(S2N)} = \boldsymbol{W}_{c,s}^{(S2N)}\boldsymbol{\zeta}_t^{(s)}, \tag{6}$$

where a 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 with the factorized transforms is given by

$$\begin{aligned}&\mathcal{Q}\left(\left\{\boldsymbol{\lambda}, \boldsymbol{W}_{1:N}^{(NAM)}, \boldsymbol{W}_{1:S}^{(SP)}, \boldsymbol{W}_c^{(S2N)}\right\}, \left\{\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{W}}_{1:N}^{(NAM)}, \hat{\boldsymbol{W}}_{1:S}^{(SP)}, \hat{\boldsymbol{W}}_c^{(S2N)}\right\}\right) \\ &\propto -\frac{1}{2}\left(\sum_{n=1}^{N}\sum_{t=1}^{T_n}\sum_{m=1}^{M}\gamma_{m,t}^{(n)}\mathcal{L}_{n,m,t}^{(NAM)} + \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{m=1}^{M}\gamma_{m,t}^{(s)}\mathcal{L}_{s,m,t}^{\prime(SP)}\right),\end{aligned} \tag{7}$$

where

$$\begin{aligned}\mathcal{L}_{s,m,t}^{\prime(SP)} &= \log|\hat{\boldsymbol{\Sigma}}_m| - \log\left|\hat{\boldsymbol{A}}_s^{(SP)}\right|^2 - \log\left|\hat{\boldsymbol{A}}_c^{(S2N)}\right|^2 \\ &+ \left(\hat{\boldsymbol{W}}_{c,s}^{(S2N)}\boldsymbol{\zeta}_t^{(s)} - \hat{\boldsymbol{\mu}}_m\right)^{\top}\hat{\boldsymbol{\Sigma}}_m^{-1}\left(\hat{\boldsymbol{W}}_{c,s}^{(S2N)}\boldsymbol{\zeta}_t^{(s)} - \hat{\boldsymbol{\mu}}_m\right).\end{aligned} \tag{8}$$

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 are used as an initial model. The speaker-dependent transforms in normal speech, $\boldsymbol{W}_s^{(SP)}$, are initialized by the traditional SAT process using only normal speech data, where the speaker-independent normal speech model is used as an initial model. In this paper, they are fixed to the initialized parameters through the proposed SAT process. They may also be updated iteratively.

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

4.3 Implementation

We have found that if both normal speech data and NAM data are used simul-

taneously for updating 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 starts to better match normal speech features rather than NAM features due to the use of a much larger amount of normal speech data compared with the amound of NAM data.

To avoid this issue, in this paper the transformed normal speech data are used for only developing the first canonical model, and then, it is further updated in SAT with 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 the conventional SAT with NAM), the model parameters are first updated using only transformed normal speech data, i.e., maximizing a part of the auxiliary function related to $\mathcal{L}_{s,m,t}^{(SP)}$ in Eq. (4) or $\mathcal{L}_{s,m,t}^{\prime(SP)}$ in Eq. (7). And then, they are further updated in the SAT process using only NAM data, i.e, maximizing only a part of the auxiliary functions related to $\mathcal{L}_{n,m,t}^{(NAM)}$, which is equivalent to the SAT process in the conventional method. In this implementation, the proposed methods are different from the conventional method only in that the initial model parameters in SAT with NAM are developed by the transformed normal speech data.

5. EXPERIMENTAL EVALUATION

5.1 Experimental Conditions

Table 1 lists training and test data. The starting acoustic model was a speaker-independent (SI) 3-state left-to-right tied-state triphone HMMs for normal speech, of 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 around 63k words (multiple pronunciations) and a 2-gram language model were used during decoding [⋆1].

Table 1 Training and test data

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 hour

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

5.2 Experimental Results

To show 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 of log-likelihoods of training utterances of NAM and normal speech through adaptive iterations in the SAT process. Within a single iteration NAM speaker-dependent CMLLR transforms and the style transforms were calculated, and then the canonical model was updated. It can be observed from this figure that during the iterative estimation, the likelihoods for normal speech data tend to increase while those for NAM data tend to decrease. We have also found that the resulting canonical model caused the degradation of NAM recognition accuracy. Therefore, the implementation described in section 4.2 was used in the following evaluation.

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. The proposed methods yield significant improvements in word accuracy (WACC) compared with the conventional method. We have found

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

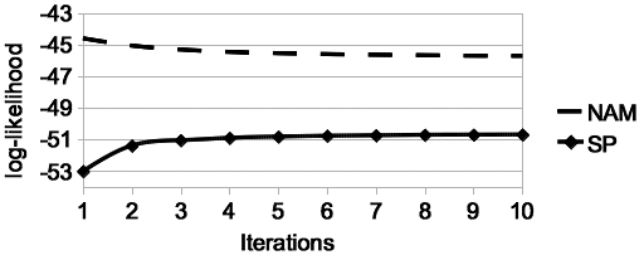

Fig. 4 Change of log-scaled likelihoods for training utterances over iterations.

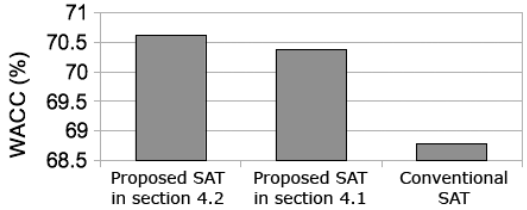

Fig. 5 Word accuracy of different methods.

that 1115 triphones models (around 1/3 of the HMM set) were not observed in NAM training data. The canonical model parameters at 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 WACC improvements shown in figure 5. Moreover, we can also observe that the use of the factorized transformation yields slight improvements in the proposed method.

6. CONCLUSIONS

In this paper, we proposed the modified speaker adaptive training (SAT) methods for building a canonical model for non-audible murmur (NAM) adaptation so as to make a larger amount of normal speech data transformed into NAM data available in the training. The experimental results demonstrated that the proposed method yields significant improvements in NAM recognition accuracy than 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 transformation in the proposed method yields a slight improvement in the performance of NAM recognition. Further investigation will be conducted on the regression tree generation of the SAT process.

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) T. Hirahara, M. Otani, S. Shimizu, T. Toda, K. Nakamura, Y. Nakajima, and K. Shikano. Silent-speech enhancement using body-conducted vocal-tract resonance signals. *Speech Communication*, Vol. 52, No. 4, pp. 301–313, 2010.
13) M.J.F. Gales. The generation and use of regression class trees for MLLR adaptation. *Technical Report*, CUED/F-INFENG/TR263, Cambridge University, 1996.