Tree-structured model selection and simulated-data adaptation for environmental and speaker robust speech recognition

Nattanun Thatphithakkul1, Boontee Kruatrachue1, Chai Wutiwiwatchai2, Sanparith Marukatat2, and Vataya Boonpiam2

1Department of Computer Engineering
King Mongkut’s Institute of Technology Ladkrabang, Bangkok, 10520, Thailand
2Human Language Technology Laboratory
National Electronics and Computer Technology Center, Pathumthani, 12120, Thailand
s6060008@kmitl.ac.th, kkboontee@kmitl.ac.th, chai@nectec.or.th, sanparith.marukatat@nectec.or.th, vataya.boonpiam@nectec.or.th

Abstract—This paper proposes the use of tree-structured model selection and simulated-data in maximum likelihood linear regression (MLLR) adaptation for environment and speaker robust speech recognition. The objective of this work is to solve major problems in robust speech recognition system, namely unknown speaker and unknown environmental noise. The proposed solution is composed of two components. The first one is based on a tree-structured model for selecting a speaker-dependent model that best matches to the input speech. The second component uses simulated-data to adapt the selected acoustic model to fit with the unknown noise. The proposed technique can thus alleviate both problems simultaneously. Experimental results show that the proposed system achieves a higher recognition rate than the system using only the input speech in adaptation and the system using a multi-conditioned acoustic model.

I. INTRODUCTION

Acoustic variation in speech recognition system can be caused by several factors e.g. speaker, environmental noise, language dialect, channel, etc. [1]. In this paper, we are interested in two main causes, namely speaker variation and environment noise variation. The basic approach to deal with these problems is to train acoustic model from noise-added speech from various speakers [2]. More elaborate techniques include model adaptation with either maximum likelihood linear regression (MLLR) [3] or maximum a posteriori (MAP) [4], parallel model combination [1], piecewise linear transformation (PLT) [5].

The major problem for every robust speech recognition system is how to handle unknown environments. Two complementary techniques, tree-structured model selection and online adaptation, can be used to tackle this problem. Tree-structured model selection consists in constructing a tree in which each node represents a combination of some known environments. An acoustic model is built for each node. Using this tree structure, an unknown environment which is similar to a combination of known environments can be better handled. This approach has been applied to select a noise-specific acoustic model [5].

The online adaptation aims at adapting the available acoustic model to the current environment. An input speech is first phone labeled given an original acoustic model. The input speech with phone labels is then used to adapt the original acoustic model and the model after adaptation is exploited in the final recognition step. Both MAP and MLLR can be used in the adaptation process. However, this technique requires a large-enough set of adaptation data in order to achieve a good recognition result. Recently Thatphithakkul et al. [6] has proposed the simulated-data adaptation process which increases the size of adaptation data by combining pre-recorded clean speeches with noise portion extracted from the current input signal. This technique allows a high gain of online-adaptation performance.

It is noted that in previous works, speaker and noise variations have often been treated separately. The objective of this work is to explore how these two variations can be handled simultaneously and efficiently. The adopted solution is based on similar idea of tree-structured model selection but used for speaker modeling instead of noise modeling propose in [5]. The speaker tree determines a speaker-dependent acoustic model which best matches to the current input signal. This tree-structured model selection can handle unknown speakers. Then we apply the simulated-data adaptation process which can solve the problem of unknown environment noise.

The proposed system is evaluated by noisy speech in 3 sets of environment. The first set contained speech in a clean environment and 9 types of noisy environments that have been trained in the system. The second set contains speech in other 2 types of noisy environments not trained in the system. Noisy speech is prepared from noise signals taken from JEIDA [7] (Japan Electronic Industry Development Association), NOISEX-92 [8] and a real noise signal collected in an exhibition in Thailand. Noise signals are added to clean speech taken from NECTEC-ATR Thai speech
corpus at various SNRs (0, 10, 15 dB). The third set contains speech signals recorded in a real environment of another exhibition in Thailand 2005. The estimated SNR of the last set is 0-5 dB.

The next section explains our proposed model. Section III describes data sets used in experiments. Experimental results are reported in Section IV. Section V concludes this paper and discusses on the future work.

II. TREE-STRUCTURED MODEL SELECTION AND SIMULATED-DATA ADAPTATION

Our proposed method of using tree-structured model selection [5] and simulated-data adaptation [6] is illustrated in Fig 1. Two principal components in this method are:

- Model selection from speaker clusters (MSSC), which functions to select the closest speaker-dependent acoustic model from a tree-structured speaker model.
- Model adaptation using simulated-data adaptation.

Section II.A describes tree-structured HMM for speaker clustering method process. Section II.B describes simulated-data adaptation process.

A. Tree-structured speaker modeling.

The tree-structured clustering method has been successfully applied for speaker adaptation [5]. In this paper, we apply the tree-structured clustering method for speaker-dependent acoustic model selection. The tree structure used in our system contains two acoustic models in each node. The first one is used in for speaker-dependent acoustic model selection, so called a classification acoustic model. Once a node is selected, the other acoustic model, called a recognition acoustic model, is used for speech recognition.

The tree-structured speaker modeling method is illustrates in Fig 2. Speeches from various speakers in various environments are collected and classification acoustic model constructed using all data. Top-down clustering is applied on the obtained acoustic model to produce the tree structure. We retrain a classification acoustic model for each cluster. Finally, the root node of the tree includes all speakers and each leaf node consists of only one speaker. Intermediate nodes in this tree contain several speakers whose speech characteristics are similar. Model selection is performed based on these speaker-dependent classification acoustic models. For the recognition acoustic model, a phoneme-based acoustic model is constructed for each noise using noise-added speech from the particular speakers. In this paper, HMM is applied to both the classification and recognition acoustic models. Using this tree structure, an unknown speaker whose sound is similar to the combination of speech from known speakers can be effectively handled.

B. Simulated-data adaptation

Since the model selection step only chooses a recognition acoustic model that best matches to the input speaker, the obtained acoustic model is yet general for every environmental noise. To enhance the system performance, online-adaptation can be conducted to make the model closest to the input environment. While the conventional online-adaptation process employs only the input signal in adaptation, the simulated-data adaptation method extends the adaptation set by adding noise extracted from the input signal to an existing set of clean speech. In this work, MLLR is conducted for adaptation. We denote S-MLLR our process of MLLR-based simulated-data adaptation. Section II.A.a describes noise extraction process. The following subsections describe in brief the simulated-data adaptation process

(a) Noise portion extraction

The silence parts are supposed to be background noise of the current input signal. An HMM is first applied to segment the input signal into speech and silence portions. The noise extraction algorithm utilizes phone-based HMMs [6].

(b) Adding background noise

Given noise portions extracted from the input signal, several issues need to be considered in adding background noise to the pre-recorded clean speech. First we concatenate noise portions extracted from the input signal. There are two noise-only regions in the input signal, at the beginning and at the end of the signal. These noise portions are duplicated and concatenated so that the duration of noise signal is equal to the duration of clean-speech being added.
Second, simulated speech for adaptation should have a similar SNR to the input speech. Let “Train-S” be a set of pre-recorded clean speech. We denote by $T_n$ and $T_s$ the current input signal and a clean speech in the Train-S set. $T_n_s$ and $T_s_s$ is the speech portion of $T_n$ and $T_s$. First, a scale factor is calculated as follows:

$$ \text{scale\_factor} = \frac{\text{EngC}}{\text{EngS}} \quad (1) $$

where EngC and EngS is the energy of $T_s_s$ and $T_n_s$ respectively. Next, the background noise, $BN$, is multiplied by the scale factor and added to $T_s$, resulting a simulated noisy-speech $T_{sn}$ as shown in (2).

$$ T_{sn} = BN*\text{scale\_factor} + T_s \quad (2) $$

### III. EXPERIMENTAL SETTING

Our task domain is isolated-word recognition using monophone-based HMMs representing 75 Thai phones. Each monophone HMM consists of 5 states and 16 Gaussian mixtures per state. 39-dimensional vectors (12 MFCC, 1 log-energy, and their first and second derivatives) are used as recognition features.

The baseline recognition acoustic model is trained by phonetically-balanced clean-speech utterances read by 16-male and 16-female speakers. The total number of training utterances is 32,000. For comparison, a multi-conditioned acoustic model [2], denoted as “MULTI” hereafter, is prepared using speech data from both clean environment and noisy environments at various SNRs (5, 10, and 15 dB). In all experiments, clean-speech data are taken from NECTEC-ATR corpus.

#### A. Noise data for training

Eight kinds of noise from JEIDA [7], including crowded street, machinery factory, railway station, large air-condition, trunk road, elevator, exhibition in a booth, and ordinary train, and one large-size car noise from NOISEX-92 [8] are conducted. All noises from JEIDA and NOISEX-92 as well as the clean speech from NECTEC-ATR are preprocessed by reducing the sampling rate to 8 kHz. Noisy speech is prepared by adding the noise from JEIDA or NOISEX-92 to the clean speech of NECTEC-ATR at various SNRs (5, 10, and 15 dB).

#### B. Noise data for testing

Two test sets, “Test-1” and “Test-2”, are used in evaluation. Test-1 contains 3,200 utterances from 640 words uttered by 5 male speakers. Two noises, a computer room from JEIDA and an exhibition (NSTDA Annual Conference S&T in Thailand) recorded over four days in March 2005, are added to clean-speech utterances at three SNR levels: 0, 10 and 15 dB. This test set represents speech with different noise from the training set.

Test-2 contains 760 utterances from 76 words uttered by 50 speakers over four days in another exhibition (ICT EXPO 2005 in Thailand). The environment is very noisy and consists of various kinds of noise. This set represents real noisy-speech with SNR ranged between 0 to 5 dB.

### IV. EXPERIMENTAL RESULTS

This section presents experimental results obtained from the proposed system including an evaluation of the MSSC technique with and without simulated-data adaptation. We also evaluate the MLLR adaptation process when the adaptation data include and exclude the input speech signal. The construction of Train-S with speakers selected by the MSSC technique is also considered. Section IV.A presents the evaluation of MSSC. The construction of Train-S is presented in Section IV.B. Section IV.C then compares our proposed system to conventional methods.

#### A. Evaluation of the MSSC model

First we evaluate the MSSC model selection technique for robust speech recognition by comparing to the baseline and the MULTI systems. Recognition accuracies obtained by test sets in three SNRs are shown in Fig 3. As expected, the MSSC, using tree-structured speaker modeling determines an acoustic model that best matches to the input speaker. Therefore, the selected model produces a higher recognition rate than the baseline and MULTI models.

![Fig. 3. Recognition accuracies of Baseline, MULTI, and MSSC evaluated by Test-1.](image)

#### B. Effects of different configurations in simulated-data adaptation

In this subsection, we are interested in the efficient way to construct the Train-S set for simulated-data adaptation. Indeed, there are two intermediate choices in building the Train-S:

- Whether the input signal should be added into the Train-S set. Input signal actually contains useful information about the acoustic characteristic of the current speaker. However, adaptation can only rely on a transcription obtained automatically. Thus the uncertainty about transcription error is also presented if we include the input speech in the Train-S set. We would like to investigate if the adaptation can still take advantage of input signal with this uncertainty. Hereafter, “supervised” adaptation refers to the adaptation
without input speech and “semi-supervised” adaptation refers to the adaptation with input speech in the Train-S set.

- Another factor which affects adaptation performance is speakers in the Train-S set. Since the acoustic model selected by the MSSC method implies the closest speaker to the input speech, the model should be adapted with speech from speakers close to the selected model. However in noisy environments, especially in low SNR, it is not always possible to correctly select speakers for the Train-S set. Therefore, we ask the question; should we adapt the selected model with all available data (limited to male-speakers 16 people, denoted as MIX) or should we use only the speech from the selected speaker cluster (denoted as “SELECT”).

In this experiment, we consider both the MULTI model and the model selected by the MSSC technique. The MULTI model adapted with simulated-data in the supervised scheme will be denoted as S-MLLR1 and the model using semi-supervised adaptation will be denoted as S-MLLR2. In analogous manner, S-MSSC1 and S-MSSC2 denote the MSSC-based selected model adapted in the supervised and semi-supervised schemes using simulated-data.

Average recognition rates for every system over all SNRs (0, 10, and 15) are shown in Fig 4. Firstly, S-MSSC outperforms S-MLLR in every case. Secondly, the semi-supervised adaptation gives higher recognition rate than the supervised one. This reflects the fact that including the input speech in adaptation is preferable. Even if the transcription of input speech might be wrong due to the automatic transcription process, the speaker characteristic in the input speech is still useful for speaker-dependent model adaptation. Lastly, MIX speakers give better performance than using only speakers in the identified speaker cluster. We believe this is due to the fact that in noisy environment, the speaker cluster selection may not be done accurately. Thus the acoustic characteristics of speakers in SELECT may be different from current speaker. On the other hand, MIX data containing all speakers should also contain the speaker close to current one. As consequence, adapting the acoustic model with SELECT only give better performance than MIX in high SNR. On average, however, using SELECT data gives lower recognition rate than using MIX. In the following, our system will function with S-MSSC using semi-supervised adaptation with MIX speakers in the Train-S set.

C. Comparison with conventional methods

In this subsection, several robust speech recognition techniques including our proposed model are experimentally compared.

Fig. 5. Comparison of Baseline+, MULTI+, MSSC+, S-MLLR and S-MSSC evaluated by Test-1 and Test-2.

The first system, called “Baseline+”, used the baseline system with online MLLR adaptation. The second system, called “MULTI+”, was a multi-conditioned acoustic model with online MLLR adaptation. The third system, called “MSSC+”, used the tree-structured speaker-dependent model selection with online MLLR adaptation. The fourth system, called “S-MLLR”, applied simulated-data adaptation to the MULTI acoustic model. The fifth system “S-MSSC” adapts the MSSC-based selected model with simulated-data. The last system is the best configuration of our propose model.

Figure 5 shows comparative results of five systems. According to results, it is obvious that our proposed method, S-MSSC, significantly outperform other conventional methods.

V. CONCLUSIONS

This paper proposed a robust speech recognition system that can deal with speaker and noise variations simultaneously. Unknown speakers and noise were handled by the tree-structured speaker-dependent model selection and online adaptation using simulated-data. Several configurations of the proposed system were investigated. Experiments showed that our proposed model achieved over 26% and 64% improvement of recognition accuracy on Test-1 (additive-noise speech) and Test-2 (real noisy speech), comparing to the conventional approach of online MLLR adaptation (Baseline+).

Future works include an evaluation of the proposed model by a larger set of speech from various environments. Further improvement of noise extraction and noise addition in simulated-data adaptation will be investigated. Moreover, the current tree structure is principally designed to handle the speaker variation. In the future work, we aim at constructing a
tree-structured model with fully support speaker as well as noise variation.

REFERENCES


