Abstract- This paper proposes a combination of simulated data adaptation and piecewise linear transformation (PLT) for robust continuous speech recognition. The original PLT selects an appropriate acoustic model using tree-structured HMMs and the acoustic model is adapted by the input speech in an unsupervised scheme. This adaptation can improve the acoustic model if the input speech is long enough and is correctly transcribed in the adaptation process. Indeed, an incorrect transcription can drastically degrade the acoustic model. Our proposed method increases the size of adaptation data by adding noise portions from the input speech to a set of pre-recorded clean speech, of which correct transcriptions are known. We investigate various configurations of the proposed method. Evaluations are performed with additive noisy continuous speech. The experimental results show that the proposed system reaches higher recognition rates than MLLR and PLT.

I. INTRODUCTION

It is commonly known that a speech recognition system trained by speech in a clean or nearly clean environment cannot achieve good performance when working in noisy environment. Research on robust speech recognition is then necessary. Gales [1] has classified the techniques of robust speech recognition into 4 approaches: 1) robust feature approach, 2) clean speech estimation approach, 3) robust model approach, and 4) combination of three previous approaches. This paper focuses on the model-based approach, which has achieved good recognition results [1]. The model-based approach aims to create or to adapt the acoustic model in specific environments. Several techniques of model adaptation have been proposed such as maximum likelihood linear regression (MLLR) [2], maximum a posteriori (MAP) [3], parallel model combination (PMC) [4], and piecewise linear transformation (PLT) [5].

In this work, we are interested in the adaptation technique of piecewise linear transformation with model selection based tree-structured cluster (MSTC) [5, 6], proposed by Zhang, Sugimura and Furui [5]. This technique is based on unsupervised acoustic model adaptation using the incoming speech. It was proven to be efficient in both the recognition accuracy and computational cost. However, a problem of the PLT is that the acoustic model may not be well adapted if the incoming speech is very short. Moreover, in the unsupervised adaptation, a wrong transcription strongly degrades the adapted model. These 2 problems can be solved by a recently proposed technique called simulated-data adaptation [7]. Indeed, the simulated-data adaptation process aims to increase the data used in adaptation by adding the background noise extracted from the current input signal to existing clean speech. Since correct transcriptions of the clean speech are known, using the simulated-data is supervised adaptation.

This paper presents a new adaptation scheme which applies simulated-data adaptation to piecewise linear transformation (called S-PLT hereafter). S-PLT combines the strength of both techniques; PLT allows selecting a good initial acoustic model while simulated-data adaptation enlarges the size of the adaptation data. S-PLT uses the MLLR adaptation technique. The proposed method is compared with other model adaptation techniques.

The proposed system was evaluated in 3 groups of environments. The first group contained a clean environment and 9 types of noisy environments that have been trained in the system. The second group contained isolated-words. The third group contained continuous-speech. Noisy speech of second and third group were prepared by adding noise signals from an exhibition in Thailand (NAC 2005) to the clean speech taken from NECTEC-ATR Thai speech corpus [8] at various SNR (0, 10, 15 dB).

We will review the PLT in the next section. Section 3 reviews the simulated-data adaptation. Section 4 describes our proposed models. Section 5 describes the data used in these works and experimental results are reported in Section 6. Section 7 concludes this paper with future works.

II. PIECEWISE LINEAR TRANSFORMATION (PLT)

The PLT method is composed of 2 main steps namely MSTC [5, 6] model selection and MLLR adaptation [2]. In the training step, a wide variety of noise data were collected and classified into noise clusters using hierarchical clustering. The root node includes all noises and all SNR conditions and each leaf node consists of only one noise at one SNR condition. Intermediate nodes in this tree contain similar noises from different environments and from different SNR. A noise-added speech HMM is constructed for each node. Using this tree structure, an unknown noise environment which is similar to combination of known environments should be handled by...
non-leaf HMM. The resulting tree-structured HMM allows representing both known and some unknown noises. In the recognition phase, the noise-cluster HMM which is best fitted to the input speech was selected and adapted by the input speech itself using the MLLR method.

III. SIMULATED DATA ADAPTATION

A. Noise portion extraction

Simulated-data adaptation begins with identifying silence parts which are supposed to be background noise of the current input signal. We assume that there are short periods of silence at the beginning and the end of the input signal. A phone-based HMM is used to segment the input signal into speech and silence portions. In this work, we use 64 HMMs of Thai phonemes including a special phoneme of silence “sil”, form a speech recognizer. Fig. 1 illustrates this HMM structure. Noise portions are the signal regions labeled with silence “sil”.

![Figure 1 HMM architecture for noise extraction.](image)

B. Background-noise addition

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 as shown in Fig. 2. These noise portions are duplicated and concatenated so that the duration of noise signal is equal to the duration of clean-speech being added.

![Figure 2 Background-noise addition.](image)

Second, simulated speech for adaptation should have a similar SNR to the input speech. However, estimation of SNR is not trivial and remains unsolved. The simulated-data adaptation relies on a simple signal-energy scaling. Let “Train-S” be a set of pre-recorded clean speech, of which correct transcriptions are known. We denote by $T_n$ and $T_s$ the current input signal and a clean speech in the Train-S set. $T_{s_n}$ and $T_{s_s}$ is the speech portion of $T_n$ and $T_s$ and $T_{s_n}$ is the silence portion of $T_n$ and $T_s$. First, a scale_factor is calculated as follows:

$$EngS = \frac{\text{sum}(\text{abs}(T_{s_n}))}{\text{length}(T_{s_n})}$$  \hspace{1cm} (1)

$$EngC = \frac{\text{sum}(\text{abs}(T_{s_s}))}{\text{length}(T_{s_s})}$$  \hspace{1cm} (2)

$$\text{scale\_factor} = \frac{EngC}{EngS}$$  \hspace{1cm} (3)

where $|T|$ denotes the energy of a signal $T$. $EngS$ and $EngC$ is respectively the energy averaged over all $T_{s_n}$ and $T_{s_s}$ extracted from $T_s$. Next, the background noise, $BN$, given by the noise-portion extraction step, is multiplied by the scale_factor and added to $T_n$, resulting in a simulated noisy-speech $T_{sn}$ as shown in Equation (4). The signal $T_{sn}$ is then included in the adaptation set.

$$T_{sn} = BN \times \text{scale\_factor} + T_s$$  \hspace{1cm} (4)

C. Adaptation-data preparation

Adaptation data set is a key element in our robust speech recognition system. Two variations of the proposed method are considered. The first one uses only the pre-recorded clean speech with known transcription. This is called supervised adaptation. The second variation includes the input speech in the adaptation set with its label transcribed automatically. This second variation is called semi-supervised adaptation.

Beside these two variations, other important parameters including the selection of speakers and lexicon in the Train-S data are required to adjust. Basically, we selected speakers and words that can be correctly recognized by our clean speech model. Moreover, selected words should cover all phones presented in the system.

IV. COMBINED SIMULATED DATA ADAPTATION AND PLT

Combined simulated-data adaptation [7] and PLT [5], called S-PLT is shown in Fig. 3.

![Figure 3 Combine simulated-data adaptation and PLT process (S-PLT) for HMM adaptation.](image)
mode. The adapted HMM is then used to recognize the incoming signal.

V. EXPERIMENTAL SETTING

Our domain is continuous speech recognition using monophone-based HMMs representing 64 Thai phones. The aim of using context-independent phone model is the flexibility to add any word in the recognition lexicon. Each monophone HMM consists of 5 states and 16 Gaussian mixture per state. 12 MFCC, 1 log-energy, and their first and second derivatives are used as recognition features. The clean model is trained by phonetically-balanced utterances read by 32 speakers. The total number of training utterances is 32,000. In all experiments, clean-speech data are taken from the NECTEC-ATR corpus [8].

D. Noise data for training

Eight kinds of noise from JEIDA, 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 are considered. All noises and the clean speech are preprocessed by resampling to 8 kHz. Noisy speech is prepared by adding the noise from JEIDA or NOISEX-92 to the clean-speech of the NECTEC-ATR corpus at various SNRs (5, 10 and 15 dB).

E. 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. Test-2 contains 1,950 utterances from 390 sentences uttered by 5 male speakers.

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.

F. Simulated-data for adaptation (Train-S set)

In order to constitute the Train-S set for model adaptation, speakers and lexicon are selected from the NECTEC-ATR corpus. For speaker selection, we limited to male-speakers with clear speech and randomly selected speech signals of four speakers are used in the experiment, denoted as “MIX” speaker.

For word selection, the selected lexicon should cover all 64 phones presented in the system. According to these criteria, two sets of train-s are prepared. The first set (Train-S1) contains 22 words and the second set (Train-S2) contains 10 sentences, both are the minimum set of isolated-words and continuous speech that cover 64 phones used in the recognizer.

VI. EXPERIMENTAL RESULTS

In this section we investigate various settings of S-PLT and compare them to other robust speech recognition techniques. Subsection A evaluates different settings of our system. Subsection B compares our proposed system to conventional methods.

A. Effects of type MLLR adaptation and type Train-S in simulated-data adaptation

In our experiments the acoustic model before adaptation can be either a clean-speech model or the model selected by the MSTC algorithm. For each model, we test adaptation in both supervised and semi-supervised modes. Semi-supervision means that the input speech is also included in the adaptation data. Table I defines four systems varying on the supervision made of adaptation and the initial acoustic model.

<table>
<thead>
<tr>
<th>Table I Definition of comparative systems</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Initial AM</td>
</tr>
<tr>
<td>Clean</td>
</tr>
<tr>
<td>Semi supervised</td>
</tr>
<tr>
<td>S-MLLR1</td>
</tr>
<tr>
<td>S-MLLR2</td>
</tr>
<tr>
<td>S-PLT1</td>
</tr>
<tr>
<td>S-PLT2</td>
</tr>
</tbody>
</table>

Fig. 4 shows the experimental results from various settings using the Train-S1 set in adaptation. According to these results, using the MSTC algorithm to initialize the acoustic model gives better accuracy than using the clean model. As expected, since the MSTC-selected model is closer to the real environment than the clean model and hence give the higher performance. For semi-supervised adaptation, results show that it rather suit for continuous speech (Test-2) than for isolated word (Test-1). This may be due to the fact that for the isolated word case, the signal is too short and no useful information can be automatically and reliably extracted. It is then better to rely only on assured transcription as in supervised adaptation. On the other hand, for continuous speech the signal is long enough to compensate the incertitude concerning the transcription thus make it useful for model adaptation.

Next, we investigate the use of continuous speech as adaptation data. For this propose, we ran the experiment on the Test-2 set with the Train-S2 adaptation data. Fig. 5 summarizes the results from this experiment. We can see that using the Train-S2 outperforms the use of Train-S1 in every case. This is due to the increasing in the size of the adaptation data. However, the computational time unavoidably increases as a drawback.
B. Comparison with conventional methods

In this subsection, several robust speech recognition techniques including ours are compared. The first one is a baseline system without any implementation for robust speech recognition. The second system, denoted as “MLLR”, uses online acoustic-model adaptation based on the well-known MLLR approach. The third system, called “MSTC”, uses model selection based on the tree-structured cluster model without any adaptation. The forth system called “PLT” uses the PLT method (see section III). The fifth and the sixth systems are S-MLLR2 and S-PLT2. The Train-S2 is used in adaptation.

Fig. 6 shows comparative results. According to the results, it is obvious that our proposed methods, S-MLLR2 and S-PLT2, outperform other conventional methods. In the case of S-MLLR2 we gain approximately 14% improvement of recognition accuracies over the conventional MLLR model. S-PLT2 achieves approximately 10% improvement of recognition accuracies over the PLT-based method. Moreover, the PLT's accuracy is lower than that of the MSTC but both systems perform worse than the S-PLT2. We believe that the continuous speech signal is not long enough, so that purely unsupervised adaptation can always improve the initial acoustic model. This underlines the benefit of using clean speech with known transcription in simulated-data adaptation.

VII. CONCLUSIONS AND FUTURE WORK

This paper proposed combined simulated-data adaptation and PLT in acoustic-model adaptation. The approach solved limitations of the conventional unsupervised MLLR adaptation in the PLT method. The adaptation data were increased by adding a noise-signal extracted from the input signal to a pre-recorded set of clean speech. Since correct transcriptions of simulated-data are given, adaptation is more effective than using only the input speech with unknown transcription. Experiments showed that our technique achieved over 10% improvement of recognition accuracy comparing to the original PLT.

Future works include an evaluation of the proposed model by a larger set of speech from various real environments. Further improvement of noise extraction and noise addition in simulated-data adaptation will also be investigated.

REFERENCES