A Simulated-Data Adaptation Technique for Robust Speech Recognition

Nattanun Thatphithakkul[1], Boontee Kruatrachue[1], Chai Wutiwiwatchai[2], Sanparith Marukatat[2], and Vataya Boonpiam[2]

[1]Department of Computer Engineering
King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand
[2]Speech Technology Section, Information Research and Development Division
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 an efficient acoustic model adaptation method based on the use of simulated-data in maximum likelihood linear regression (MLLR) adaptation for robust speech recognition. Online MLLR adaptation is an unsupervised process which requires an input speech with phone labels transcribed automatically. Instead of using only the input signal in adaptation, our proposed simulated data method increases the size of adaptation data by adding noise portions extracted from the input speech to a set of pre-recorded clean speech, whose correct transcriptions are known. Various configurations of the proposed method are explored. Evaluations are performed with both additive and real noisy speech. The experimental results show that the proposed system achieves higher recognition rate than the system using only the input speech in adaptation and the system using a multi-conditioned acoustic model.

Index Terms: robust speech recognition, MLLR, online-adaptation

1. 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. 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. Research works on the model-based approach have been extensively carried out. Figure 1 illustrates a normal recognition process with online-adaptation. 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 maximum a posteriori (MAP) adaptation [2] and maximum likelihood linear regression (MLLR) [3], and [4] are efficient adaptation algorithms.

The model presented in Figure 1, however, has two major limitations. First, the MLLR or MAP requires a large-enough set of adaptation data in order to achieve a good recognition result. In real world applications, users often input a very short sentence or, worst, only an isolated-word, which limits the improvement of adaptation. Second, online-adaptation is unsupervised adaptation, i.e. it uses phone-labels transcribed automatically by the original acoustic model. Given the original acoustic model, which may incorrectly transcribe the input speech, the adapted model cannot yield a satisfied result.

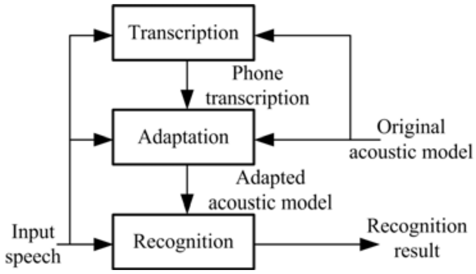

Figure 1 *A recognition process with online-adaptation.*

This paper proposes a novel approach of simulated-data adaptation, which resolves two limitations mentioned above. The simulated-data adaptation approach increases adaptation data by adding background noise extracted from the input signal to a pre-recorded set of clean speech, whose correct transcriptions are known. This process not only increases the size of adaptation set, but also reduces the problem of using incorrect transcriptions in adaptation. The MLLR algorithm performs faster and better than the MAP when the adaptation set is small, whereas the MAP becomes asymptotically more accurate than the MLLR when the size of adaptation set increases [5]. Since one of our concerns is real-time processing, the size of the adaptation data cannot be very large. In this condition, we choose only the MLLR adaptation in experiments.

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 (Japan Electronic Industry Development Association) [6] and a real noise signal collected in an exhibition in Thailand (NAC 2005). Noise signals are added to clean speech taken from NECTEC-ATR Thai speech corpus [7] at various SNRs (0, 10, 15 dB). The third set contains speech signals recorded in a real environment of another exhibition in Thailand (ICT-EXPO 2005). The estimated SNR of the last set is 0-5 dB.

The next section explains our proposed model. Section 3 describes data sets used in experiments. Experimental results are reported in Section 4. Section 5 concludes this paper and discusses on the future work.

2. Simulated-Data Adaptation

Our proposed method of using simulated-data in MLLR [4] adaptation, denoted as "S-MLLR", is illustrates in Figure 2. While the conventional process employs only an input signal in acoustic model adaptation, the S-MLLR method extends the adaptation set by using simulated-data created by adding noise extracted from the input signal to an existing set of clean speech. As described in the introduction, simulated-data adaptation overcomes problems of data sparseness in adaptation and unknown label of the input speech. Two issues are considered in the proposed method. The first issue is how to accurately extract noise portions from the input speech. Section 2.1 describes our noise extraction process. Once having extracted the noise portion, the second issue is how to add the noise signal to a given set of clean speech. We explain the process of adding noise in Section 2.2.

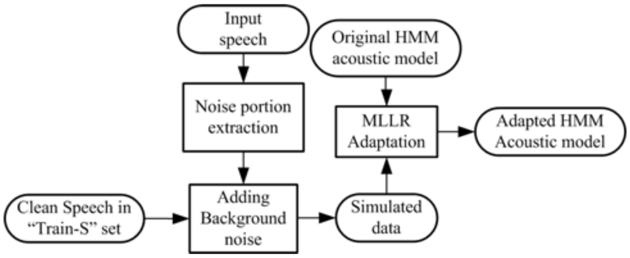

Figure 2 Simulated-data MLLR process (S-MLLR) for HMM adaptation.

2.1. Noise portion extraction

Simulated-data adaptation begins with identifying silence parts in the input signal. The silence parts are supposed to be background noise of the current input signal. For our task of isolated-word recognition, we assume that there are short periods of silence at the beginning and the end of the input signal. A hidden Markov model (HMM) is used to segment the input signal into speech and silence portions. Two noise extraction algorithms are evaluated in this paper. The first algorithm utilizes phone-based HMMs, where 64 HMMs of Thai phonemes including a special phoneme of silence "sil", as shown in Table 1, form an isolated-word recognizer. Figure 3(a) illustrates this HMM structure. The second noise extraction algorithm is based on speech/non-speech detection. Two states HMM, symbolized with speech and silence, are included in the module as shown in Figure 3(b). In both algorithms, noise portions are the signal regions labeled with silence "sil".

Table 1. *64 Thai phonemes.*

Type	IPA symbol
Initial consonant	p, t, c, k, ʔ, p^h, t^h, c^h, k^h, h, b, d, m, n, ŋ, l, r, f, s, h, w, j, pr, pl, p^hr, p^hl, tr, t^hr, kr, kl, kw, k^hr, k^hl, k^hw, fr
Vowel	i, iː, ɨ, ɨː, u, uː, e, eː, ɜ, ɜː, o, oː, æ, æː, a, aː, ɔː, iːa, ɨːa, uːa
Final consonant	p', t', k', m', n', ŋ', s', w', j'
Silence	sil

In both algorithms, HMMs are composed of 16 Gaussian mixtures per state and were trained by the Baum–Welch algorithm. It is noted that the former algorithm gives better noise-region labeling performance with a drawback of computational demand comparing to the latter algorithm.

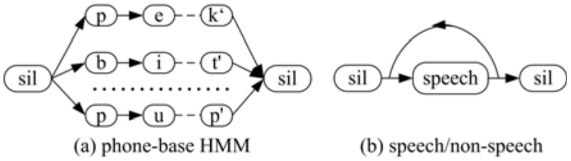

Figure 3 Two HMM architectures for noise extraction.

2.2. 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 as shown in Figure 4. These noise portions are duplicated and concatenated so that the duration of noise signal is equal to the duration of clean-speech being added. It is noted that simply concatenating noise portions causes an unusual spectral change. However, in this paper, we discard spectral smoothing in order to save processing time.

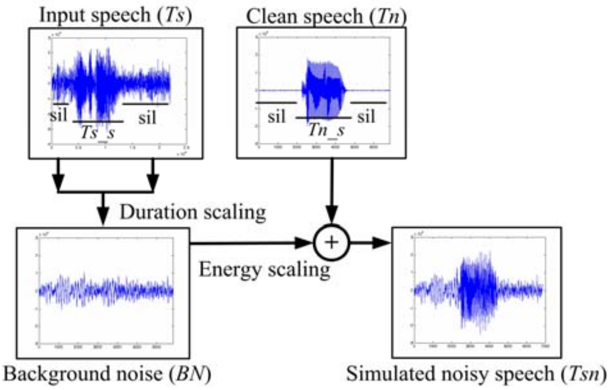

Figure 4 *Adding background noise.*

Second, simulated speech for adaptation should have a similar SNR to the input speech. However, estimation of SNR is not trivial and remains unsolved. In this work, we propose a simple way of signal-energy scaling. Let "Train-S" be a set of pre-recorded clean speech, of which correct transcriptions are known. We denote by *Tn* and *Ts* the current input signal and a clean speech in the Train-S set. *Tn_s* and *Ts_s* is the speech portion of *Tn* and *Ts* and *Tn_sil* and *Ts_sil* is the silence portion of *Tn* and *Ts*. First, a *scale_factor* is calculated as follows:

$$EngC = sum(abs(Ts\_s))/length(Ts\_s) \qquad (1)$$

$$EngS = sum(abs(Tn\_s))/length(Tn\_s) \qquad (2)$$

$$scale\_factor = EngC/EngS \qquad (3)$$

where *EngC* and *EngS* is the energy of *Ts_s* and *Tn_s* respectively. Next, the background noise, *BN*, is multiplied by the *scale_factor* and added to *Ts*, resulting a simulated noisy-speech *Tsn* as shown in Equation 4.

$$Tsn = BN*scale\_factor + Ts \qquad (4)$$

3. Experimental Setting

Our domain is isolated-word recognition using monophone-based HMMs representing 64 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 acoustic model (clean-speech model) is trained by phonetically-balanced 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 [8], denoted as “MULTI” hereafter, is trained by 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 [7].

3.1. Noise data for training

Eight kinds of noise from JEIDA [6], 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 [9] 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).

3.2. Noise data for testing

Two test sets, “Test-1” and “Test-2”, are used in evaluation. Test-1 contains 3,200 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 utterances covering 76 Thai-province names recorded from 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.

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

In order to constitute the Train-S set for model adaptation, several criteria are used to select speakers and lexical words from the NECTEC-ATR corpus. For speaker selection, we limited to male-speakers with clear speech. Four speakers, denoted as “M1” to “M4”, are selected.

For word selection, two criteria are considered. First, these words should be correctly recognized by our clean-speech model. Second, words should cover all 64 phones presented in the system. According to these criteria, 22 words out of 76 Thai-province names are chosen.

4. Experimental Results

Experiments are organized as follows. First, several parameters in the adaptation process are optimized. Among various parameters, we have found that the number of speakers and the size of adaptation data were influential. Section 4.1 gives the detail of optimization of these parameters. Given optimized parameters, Section 4.2 then compares our proposed system to conventional methods.

4.1. Effects of different speakers and data size in simulated-data adaptation

In the case that speech signals from only one speaker are included in the simulated-data set, increasing the size of adaptation data tends to produce a speaker-dependent acoustic model. Using the speaker-specific acoustic model may reduce recognition accuracy when evaluating with speech from various speakers. Therefore, in this subsection, five experiments on S-MLLR are performed to explore this phenomenon. Each of the first four experiments uses speech of only one speaker (M1 to M4). Randomly selected speech signals of M1 to M4 speakers are used in the last experiment, denoted as “MIX” case. Both noise extraction and MLLR adaptation in S-MLLR utilize phone-based HMMs trained by multi-conditioned data, i.e. speech data from clean and various noisy environments.

Figure 5 plots accuracies averaging on Test-1 and Test-2. According to results, the accuracy increases as a function of data size. We conclude that the phenomenon of speaker-mismatching is not significant even when a large number of speakers is evaluated (50 speakers in Test-2). The use of MIX case, where the adaptation set contains speech from various speakers, gives obviously better performance than the use of one specific-speaker model. This indicates that the acoustic model should be adapted with its speaker-independent property maintained.

Figure 6 plots processing time of the MIX case. All experiments are performed on an Intel Pentium IV 3.2 GHz CPU with 2 GB RAM. The graph shows that increasing the size of adaptation data yields higher processing time. An optimal number of words in the Train-S set we choose for the rest experiments is 22. We recall that these 22 words are the minimum set of words that covers all 64 phones.

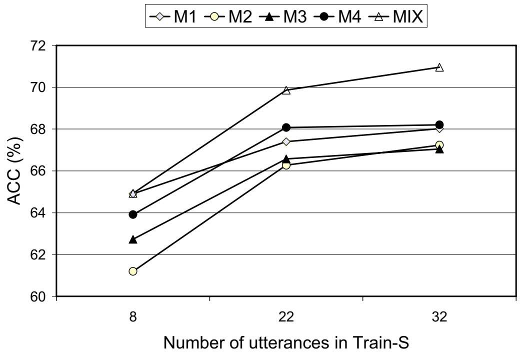

Figure 5 *Recognition accuracy of S-MLLR with different data sizes and selected speakers in Train-S.*

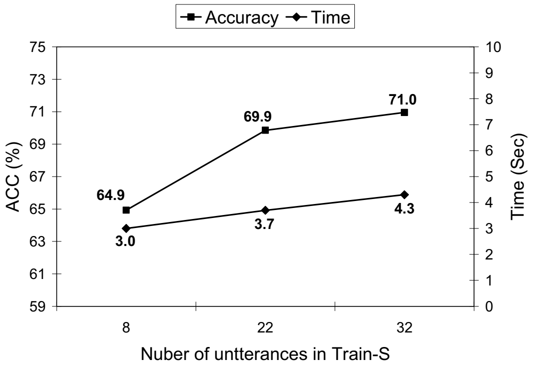

Figure 6 *Recognition accuracy and processing time of S-MLLR with different data sizes of Train-S.*

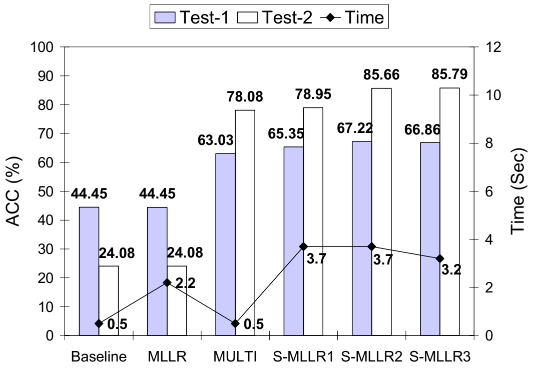

Figure 7 Comparison of Baseline, MLLR, MULTI, and S-MLLR systems evaluated by Test-1 and Test-2.

4.2. Comparison with conventional methods

In this subsection, several robust speech recognition techniques including our proposed model are experimentally compared. The first system is a baseline system without any implementation for robust speech recognition. The second system, denoted as "MLLR", exploits a conventional technique of online acoustic-model adaptation using MLLR as illustrated in Figure 1. The third system, called "MULTI", used a multi-conditioned acoustic model without any adaptation. The rest three systems are based on our proposed S-MLLR method. The forth system, called S-MLLR1, utilizes phone-based HMMs for noise extraction and online MLLR adaptation. The phone-based HMM is trained by clean-speech. The fifth system, S-MLLR2, is similar to the S-MLLR1 system except that the phone-based HMM was multi-conditioned. The last system, S-MLLR3, replaces the phone-based noise extraction module by a speech/non-speech detection module described in Section 2.1. The Train-S set contains 22-word utterances from M1 to M4 training speakers.

Figure 7 shows comparative results of five systems evaluated by Test-1 and Test-2. According to results, it is obvious that our proposed methods of S-MLLR outperform other conventional methods. Comparing among variations of S-MLLR, S-MLLR2 gives the highest accuracy but takes the longest processing time. S-MLLR3 is the fastest with the accuracy between the other two systems. We conclude that the phone-based HMM trained by multi-conditioned data in S-MLLR2 gives the best noise extraction result and hence causes the highest recognition accuracy. The speech/non-speech detection module in S-MLLR3 is much simpler than the phone-based HMM but achieves comparable performance.

5. Conclusions

This paper proposed a new approach of using simulated-data in MLLR acoustic-model adaptation. The approach solved limitations of the conventional online MLLR adaptation. The adaptation data was increased by conducting simulated-data created by adding a noise signal extracted from 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 proposed model achieved over 20% improvement of recognition accuracy comparing to the conventional approach of online MLLR adaptation.

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 be investigated. Another interesting issue is that selection of clean speech from different speakers should affect the recognition result. Even if we have found that maintaining speaker-independency in adaptation gives good recognition result, selecting a set of speakers that best matches to the input speech might give better performance. This issue will also be explored.

6. References

[1] Gales, M.J.F., "Model-based techniques for noise robust speech recognition", PhD. Thesis, University of Cambridge, 1995.

[2] Gauvain, J.L., and Lee, C.H., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", IEEE Trans. Speech Audio Proc., 2:291–298, 1994.

[3] Leggetter, C.J. and Woodland, P.C., "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs", Comput. Speech Lang., 9:171–186, 1995.

[4] Gales, M.J.F. and Woodland, P.C., "Mean and variance adaptation within the MLLR framework", Comput. Speech Lang., 10(3):249–264, 1996.

[5] Zhao, Y., Wang, L., Chu, M., Soong, F.K. and Cao, Z., "Refining phoneme segmentations using speaker-adaptive context dependent boundary models", Proc. of INTERSPEECH 2005, pp.2557-2560, 2005.

[6] http://www.milab.is.tsukuba.ac.jp/corpus/noise db.html

[7] Kasuriya, S., Sornlertlamvanich, V., Cotsomrong, P., Jitsuhiro, T., Kikui, G. and Sagisaka, Y., "NECTEC-ATR Thai speech corpus", Proc. of Oriental COCOSDA 2003, pp.105-111, 2003.

[8] Nakamura, S., Yamamoto, K., Takeda, K., Kuroiwa, S., Kitaoka, N., Yamada, T., Mizumachi, M., Nishiura, T., Fujimoto, M., Saso, A., Endo, T., "Data collection and evaluation of AURORA-2 JAPANESE corpus", Proc. of ASRU 2003, pp.619-623, 2003.

[9] http://www.speech.cs.cmu.edu/ comp. speech/ Section1/ Data/noisex.html