A study of HMM-based automatic segmentations for Thai continuous speech recognition system

Pongthai Tarsaku Supphanat Kanokphara
Information R&D Division,
National Electronics and Computer Technology Center (NECTEC)
112 Paholyothin Rd., Klong 1, Klong Luang, Pathumthani 12120 Thailand
e-mail : pongthai@nectec.or.th, supphanat_k@notes.nectec.or.th

Abstract

Speech segmentations have been widely using in many speech applications. In speech synthesis, the quality of produced speech depends on the accuracy of labeled acoustic inventory. In speech recognition, segmented utterances according to the labels are usually used as a starting point for training speech models. The segmentation is often manually encoded which is time-consumption process and has low precision and consistency in some parts of speech. Therefore, the new better technique for speech segmentation using HMM is proposed.

In this framework, the effects of manual and automatic segmentation are examined by using the output of final application, the word accuracy of speech recognition. From the experiment, manual segmentation has only 0.59 % better than mono-phone automatic segmentation. The result convinces that tri-phone automatic segmentation will give a better result in the future work.

1 Introduction

Nowadays, HMM takes parts deeply in the field of speech technologies. This is because the properties of HMM which can normalize speech signal's time variation and represent the speech signal statistically. In the HMM training process, the Baum-Welch method (Rabiner, 1989) is frequently used to optimize the speech model. However, one of the weak points of Baum-Welch optimization is that the estimated model might correspond only to a local likelihood maximum, not the global one.

One way to solve this problem is to find a good initial point for model parameters. As a result, segmented utterances corresponding to the labels are needed as a starting point for training HMM parameters in order to make local maximum likelihood equal or as close as possible to the global one.

The segmented utterances are often be done by expert labelers. By this fashion, time for implementing speech applications is vastly spent on the data preparation process. Moreover, some parts of the speech are too complicated to be classified by a human listener, and those parts must be segmented by the assumption such as segmentations between nasal and nasal, vowel and approximant, etc.

HMM-based phonetic recognition is very fast, precise and consistent compared with the human. However, as human has ability to solve fuzzy problem better than computer can do, the superiority between manual and automatic segmentation is still unclear for a speech recognition task.

This paper is organized as follows. This section includes introduction and the reason of using word recognition result as the evaluation method. Section 2 is about automatic segmentation system. Section 3 describes the experimental method of this paper. Section 4 shows the result of the experiment. The conclusion and future work is discussed in Section 5.

1.1 Continuous Speech Recognition as an Evaluation Method

Most of the studies of segmentation usually assume that hand label is definitely correct, and they (Nefti, 2001) evaluate the segmentation system by comparing the result from the automatic segmentation with the manual one. The goal of this study is, however, not to make the automatic segmentation system close to the manual segmentation but to prove that computer can outperform human in this task.

As stated earlier, by using Baum-Welch training, an efficient model can be obtained if

the initial parameters are close to the global maximum point. Therefore, a better-quality segmentation method would lead to the more efficient model resulting in higher accuracy recognition rate finally.

Even Makashay et al. (Makashay, 2000) have shown that their automatic segmentation is better than manual one by listening tests (scoring the quality of speech output from TTS), this might be biased as classification of each people is different. By this reason, evaluation by recognition accuracy, should, contrarily, give the fair comparison of the segmentations.

2 Automatic Segmentation

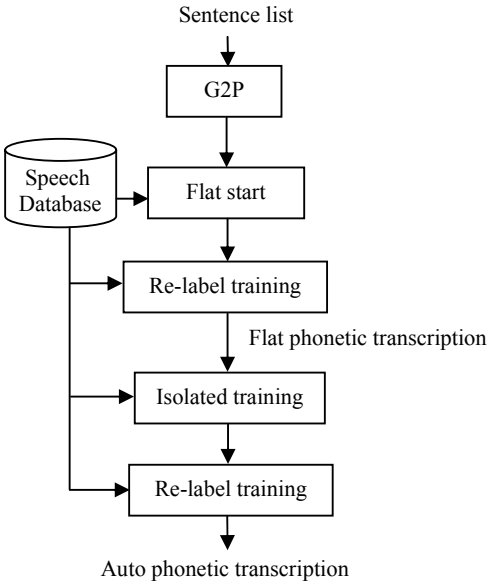

Figure 1: Automatic Segmentation system.

The automatic segmentation here uses HTK (Young, 2000) as a base system. Automatic segmentation requires sentence list and corresponding speech database as inputs. The system starts from the automatic Thai Grapheme-to-Phoneme (G2P) developed by NECTEC (Tarsaku, 2001). This process generates phonetic transcriptions and dictionary as inputs for flat start. The flat start constructs an acoustic model for re-label training. The re-label training updates phonetic transcriptions. These transcriptions are then the inputs for isolated training. The isolated training produces the acoustic model for the second re-label training. The second re-label training generates auto phonetic transcriptions. The procedure is shown in Figure 1.

The flat start is a model starting technique of HTK in the case that has no time-alignment phonetic transcription. This technique starts from calculating global means and variances of all speech parameters, using those parameters as the beginning point of each phone model and retraining those models by using Baum-Welch to obtain the optimum model. This process is used to create initial acoustic model for first re-label training.

The isolated training is another training technique of HTK. Each phone is trained separately according to the transcriptions. By doing this, new segmentations are constrained and cannot go beyond sub-word boundary. Therefore, new segmentations have less error due to unsaturated model. The isolated training can be separated into two parts: hard and soft boundary segmentation. The process starts from hard to soft boundary training respectively in order to reduce error from unsaturated model. Hard boundary training segments speech database separately while soft boundary training allow some probabilistic overlap around the boundaries.

2.1 Re-label training

As there is still some error from G2P, there are three more processes applied during re-label training: pronunciation correction, short pause insertion and phonetic alignment. Force Viterbi algorithm is used for these processes. Re-label training can update transcription during training in order to obtain maximum likelihood transcription. Figure 2 shows re-label training process.

Pronunciation Correction: As the pronunciation from G2P still has some error due to the complexity in Thai pronunciation variation, the pronunciation for training may deteriorate the quality of the training model. However, the dictionary generated from G2P limits the variation in pronunciation of each word. By using Viterbi algorithm, the correct pronunciation can be selected from those pronunciation candidates.

Short Pause Insertion: In the database, there are many long words that the speaker cannot say it without breathing. Therefore, there

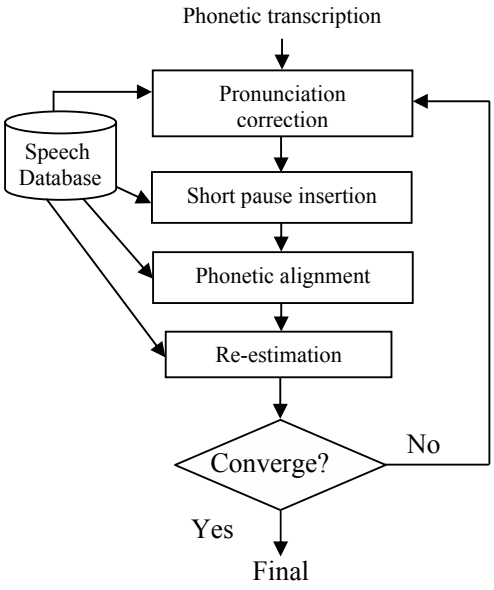

Figure 2: Re-label training.

might be a short pause during recording process. Automatic short pause insertion is built for supporting this error. For the short pause insertion, there is an algorithm searching the possible point of short pause in a word (usually after syllable). Then, the force Viterbi is employed to find the best pronunciation.

Automatic Phonetic Alignment: After pronunciation correction and short pause insert processes, the correct phonetic transcriptions are generated. Then with this phonetic transcriptions and force Viterbi, the time alignment transcriptions are created.

Pronunciation correction, short pause insertion and phonetic alignment use input acoustic model to update input transcriptions for re-estimation process. The re-estimation process then builds acoustic model as the input of those three processes. This process repeats until the log probability of update model is less than the last one.

3 Experimental Methods

The experiment procedure starts from data preparation, wave to MFCC conversion, making topology prototype, label and dictionary construction (in HTK format), training models, and testing finally.

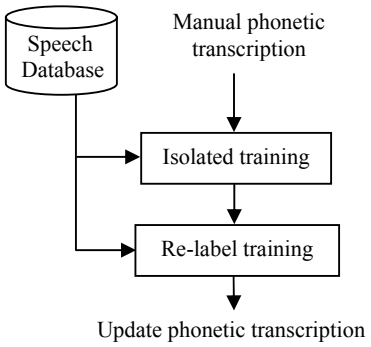

Figure 3: Base line system

In this experiment, three phonetic transcriptions are compared. Flat transcriptions are the output from first re-label training and auto transcriptions are the result from second re-label training as shown in Figure 1. Both transcriptions are evaluated with update transcriptions trained by isolated training and re-label training. The input phonetic transcriptions for this training are manually edited. Figure 3 demonstrates this base line system.

For decoding, a back-off bi-gram language model is constructed and Viterbi algorithm is applied for speech recognition process.

3.1 Database

There are 1,246 utterances for training and 140 utterances for testing, which are recorded by one speaker. The database consists of 3,097 words and 76 sub-word units including silence and short pause unit.

The update transcriptions making process consists of taking the output from the aligner program and manually correcting it. The aligner segmentation for each utterance was examined by using Wavesurfer 1.0.4 (Sjolander, 2000). There are only 2 expert labelers for the correction process in order to preserve the consistency. Complicated points in transcription and boundary alignment are discussed and adopted during the process.

The language model is constructed from 1,246 sentences according to the utterances. Back-off bi-gram's perplexity is 73.68 and entropy is 6.20.

Speech utterances are parameterized into 12 dimensional vectors, energy, and their delta and acceleration (39 length front-end parameters).

Output transcription	Training log probability	% Correct	% Accuracy
Flat	-58.73	73.66	71.18
Update	-58.63	74.39	72.15
Auto	-58.71	74.01	71.56

Table 1. The recognition result form different initial model

Iteration	Flat	Update	Auto
0	-59.58	-58.63	-58.71
1	-58.77	X	X
2	-58.73	X	X

Table 2. The training log probability in each iteration in re-labeling re-estimation process

4 Experimental Results

At this point, three transcriptions are compared. The training probability and percent correct and accuracy of test set according to the particular models are presented. These three properties show the performance of the models in various aspects. The training probability implies the recognition score with training set while others two properties show the recognition score with testing set. These data show precision and robustness of the acoustic model.

4.1 Auto vs. flat transcriptions

Even though flat start is one of the good techniques for initialization without time-alignment transcriptions, model that is trained by using segmented labels proves to produce better results in terms of recognition rate. By observing Table 1, Auto confirms better results of recognition rate than Flat. This can obviously be seen by the different between Flat and Update results.

Table 2 shows that Auto and Update requires no re-label training. This means starting points of both systems are close to the global optimum point.

4.2 Update and Auto transcriptions

From Table 1, Update shows only slightly better results than Auto. This shows that there are some complicated points of speech that cannot be segmented by using only mono-phone automatic segmentation. By observing Table 2, update and auto transcriptions require no re-label training. Therefore, input transcriptions are the output transcription.

5 Conclusion

From this experiment, it can be concluded that manual segmentation is better than mono-phone automatic segmentation. This is because mono-phone cannot reduce error from phone articulation. With this reason, tri-phone automatic segmentation will be experimented in the future work.

References

L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recogition", In Proc. IEEE, 77 volume 2, pp 257 – 286, 1989

S. Young D. Kershaw, J. Odell, D. Ollason, V. Valchev, P. Woodland, "The HTK book", http://htk.eng.cam.ac.uk/docs/docs.shtml, 2000

S. Nefti., O. Boeffard, "Acoustical and Topological Experiments for an HMM-based Speech Segmentation System", In Proc. *Eurospeech*, pp. 1711 – 1714, 2001

M.J. Makashay, C.W. Wightman, A.K. Syrdal, A. Conkie, "Perceptual Evaluation of Automatic Segmentation in Text-to-speech Synthesis", In Proc. *ICSLP*, volume 2, pp. 431–434, 2000

K. Sjölander, J. Beskow, "Wavesurfer – An Open Source Speech Tool" In Proc. *ICSLP*, volume 4, pp. 464 – 467, 2000

P. Tarsaku, V. Sornlertlamvanich, R. Thongprasirt, "Thai Grapheme-to-Phoneme using Probabilistic GLR Parser" In Proc. Eurospeech, volume 2, pp. 1057 – 1060, 2001