Improving ASR for Continuous Thai Words Using ANN/HMM

Maleerat Sodanil [1], Supot Nitsuwat[1] , Choochart Haruechaiyasak[2]

[1]Department of Information Technology
King Mongkut's University of Technology North Bangkok (KMUTNB)
1518 Pibulsongkarm Rd., Bangsue, Bangkok 10800

[2]Human Language Technology Laboratory (HLT)
National Electronics and Computer Technology Center (NECTEC)
Thailand Science Park, Pathumthani 12120, Thailand

msn@ kmutnb.ac.th
sns@kmutnb.ac.th
choochart.haruechaiyasak@nectec.or.th

Abstract: The baseline system of an automatic speech recognition normally uses Mel-Frequency Cepstral Coefficients (MFCC) as feature vectors. However, for tonal language like Thai, tone information is one of the important features which can be used to improve the accuracy of recognition. This paper proposes a method for building an acoustic model for Thai-ASR using a combination of MFCC and tone information as an input feature vector. In addition, we apply Artificial Neural Network (ANN) multilayer perceptrons to estimate the posterior probabilities of a class model given a sequence of observation input. The performance of the ANN approach is compared with the Gaussian Mixture Model (GMM) used in the Hidden Markov Model Toolkit (HTK). The experiments were carried out with 2-grams and 3-grams of language model. The training and test data sets were prepared from reading speech of ten Aesop's stories from 5 male and 5 female speakers. The results showed that the proposed method can be used to improve the performance of Thai-ASR in term of reducing word error rate.

1 Introduction

The challenge in Automatic speech Recognition (ASR) is how to improve the accuracy of speech recognition in term of performance of the algorithm. There are three main parts of ASR, the first one is feature extraction that extracts distinguished feature of speech utterance, the second is training model which is typically based on the Hidden Markov Model (HMM) framework and the third is decoder which finds the best probabilistic match between speech utterance and text transcription. For tonal language like Mandarin or Thai in which tone is important for specifying the meaning of speech utterance, therefore, tone information could be considered in the system in order to improve the accuracy of speech recognition. There have been some researches proposing tone recognition or classification [SM1999, SY1995, NB2002] for improving the accuracy of speech recognition [CT2006]. Although well-known Mel-frequency cepstral coefficients (MFCC) features and HMM are widely used as feature vectors, there are some concern about testing, combining and adapting them to improve the accuracy or performance of the system which may not depend on speakers or languages [XM2006, PA2008].

The HMM is a very powerful statistical method for characterizing the observed data samples of a discrete time series. In HMM, the states are not directly visible, but variables influenced by the state are visible. Each state has a probability distribution over the possible output observation. The state transitions are also probabilistic in nature. The complete HMM model is denoted as $\lambda = (A,B,\pi)$. The HMM training procedure tries to estimate the value of state transition probability distribution (A), observation symbol probability density or emission probability (B), and initial state distribution (π). The emission probability distribution function (PDF) estimates the probability with which a given observation has been generated. However, the standard HMM based on maximum likelihood criteria (ML) has some weakness caused by several assumptions which reduce discriminative power in classification. PDF is mostly computed by gaussian mixture distribution function as baseline system in order to reduce the number of trainable parameters and lower the computational costs [DA1994]**.** However, neural network based on the conventional forward-backward algorithm has also been used to estimate the posterior probability of the state distribution given an observation sequence of speech utterance [YM1997].

Recently, Reynolds and Antoniou [TC2003] investigated the use of a layered modular/ensemble neural network architecture for acoustic modeling. This architecture decomposes the task of acoustic modeling by phone. Pavelka and Ekštein [TK2009] used the hybrid of neural network (NN) to estimate the state emission probabilities which reduce word error rate compared with GMM/HMM. These probabilities are used as the HMM state emission probabilities to perform the Viterbi decoding to find the result path of word which the system can be recognized.

Accoding to the advantage of tone information and the HMM framework which used extensively in speech recognition to model the temporal information in speech and neural network which more powerful tool for a clssification tasks due to their discriminant nature of speech manner [AT1989]. In this paper, the combination of tone and MFCC was used as input of acoustic model. The neural network multilayer perceptron is used to estimate the state emission probabilities for all classes which corresponding to phonetic units.

The remaining of this paper is organized as follows. The feature extractions will be introduced in Section 2. In Section 3, an acoustic model will be described with two methods of GMM/HMM and ANN/HMM. The experimental results and conclusion were summarized in Section 4 and Section 5, respectively.

2 Features Extraction

The objective of features extraction is to extract characteristics from the speech signal that are unique to each word which will be used to differentiate between a wide set of distinct words. In this paper, the combination of MFCC and tone features was used as an input of acoustic model.

2.1 MFCC Features

Mel-frequency cepstral coefficients (MFCC) is considered as the standard method for feature extraction in speech recognition systems. The MFCC computational starts with pre-emphasis. Then the continuous time signal (speech) is sampled at discrete time points to form a sample data signal representing the continuous time signal. The samples are quantized to produce a digital signal. Next step is framing using hamming window. In this paper, the input speech signal is segmented into frames of 25 ms of frame size with optional overlap of 10 ms. Each frame has to be multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame. The discrete Fourier transform (DFT) is normally computed via the fast fourier transform (FFT) algorithm to evaluate the frequency spectrum of speech. FFT converts each frame of N samples from the time domain into the frequency domain and obtain the magnitude frequency response of each frame. Then the magnitude coefficients of the fourier transform for the speech segment is binned by correlating it with each triangular filter in the filterbank. In this paper, 24 filterbanks are used. The logarithm state simply converts the multiplication of the magnitude of the fourier transform into addition such as log energy within a frame. The final procedure for the mel frequency cepstral coefficients computation consists of performing the inverse of DFT on the logarithm of the magnitude of the mel filterbank output referred to as signal's mel cepstrum. In our experiments, 13 MFCC features plus deltas and double-deltas parameters are extracted using HTK.

2.2 Tone Features

For Thai language, the syllable consists of three parts: initial consonant, vowel and final consonant respectively. Each syllable has its tone. The fundamental frequency (F_0) or pitch can be extracted from voiced part of the time unit in the utterance. Normally, in vowel position of syllable. Therefore, the F_0 needs to be interpolated in unvoiced regions to avoid variance problems in recognition using a smoothed log-pitch estimate and its two temporal derivatives [XM2006]. In this paper, F0 is extracted and smoothed , then combined to the standard MFCC to be used as an input of acoustic model. The average magnitude difference function is used instread of autocorrelation function to extract pitch period. It computes the difference between the signal and time shifted version of itself. The average magnitude difference function [MH1974] is defined as :

$$AMDF(\tau) = \frac{1}{N}\sum_{n=0}^{N-1-\tau} |x(n) - x(n-\tau)| \quad (1)$$

Where $x(n)$ are the samples of analyzed speech frame
$x(n+\tau)$ are the samples time shifted τ seconds and N is the frame size.
The smoothing using the moving average smoothing is used as the following equation.

$$\widehat{F_n} = \frac{1}{N}\sum_{i=n-N/2}^{n+N/2} F_i \quad (2)$$

where F_i is the order i of F_0, $\widehat{F_n}$ is the smoothed F_0 of frame n, and N is the frame size.
In order to solve the end-effect problem, a simple first order differences at the start and end of the speech was used as following

$$delta_n = \begin{cases} \frac{f_{n+\theta} - f_{n-\theta}}{2\theta}, & \theta < n < N-\theta \\ f_{n+1} - f_n, & n < \theta \\ f_n - f_{n-1}, & n \geq N-\theta \end{cases} \quad (3)$$

where $delta_n$ is a delta coefficient at time n, θ is the internal distanace between two F_0, f_n is a smoothed F_0 value at time frame n and N is to total frame. The total of tone feature equal 3 feature vectors.

3 Acoustic Models

Acoustic modeling plays a critical role in improving accuracy of any speech recognition system. For the given acoustic observation O = $O_1 O_2 ... O_n$ the goal of speech recognition is to find out the corresponding word sequence that has the maximum posterior probability P(W|O) as expressed by Eq. (4).

$$\widehat{W} = argmax_w\ P(W|O) = \ argmax_w\ \frac{P(W)P(O|W)}{P(O)} \tag{4}$$

Since the maximization of Eq. (4) is carried out with the observation O fixed, the above maximization is equivalent to maximization of the following equation:

$$\widehat{W} = argmax_w P(W)P(O|W) \tag{5}$$

where P(O| W) is acoustic models and P(W) is language models, that can truly reflect the spoken language to be recognized. In this paper, an acoustic model is considered to improve the accuracy of speech recognition using neural network multilayer perceptron. The combination of MFCC standard feature vectors and tone features are given as input to ANN which will be described. To measures speech recognition error and evaluate the performance of the system. The word recognition error rate is widely used as one of the most important measures. The *Word Error Rate* is defined as:

$$WER = \frac{N - S - D - I}{N} \times 100\% \tag{6}$$

where N is the total number of words. S, D and I are number of word substitutions, deletions and insertions respectively.

3.1 GMM/HMM

A gaussian mixture model (GMM) which parameterized by a mean and a variance often modeled to estimate the emission probability density of an HMM framework. In the hidden markov toolkit (HTK), the parameter estimation was done by a flat start embedded training which required the phonetic transcriptions of training utterances to be available. HTK allows each observation vector at time t to be split into a number of S independent data streams o_{st}. The formula for computing $b_j(o_t)$ is then defined as

$$b_j(o_t) = \left[\prod_{s=1}^{S}\sum_{m=1}^{M_s} b_{jm}(o_{st}; \mu_{jm}, \Sigma_{jm})\right]^{\gamma_s} \tag{7}$$

where M_s is the number of mixture components in stream s, c_{jm} is the weight of the m'th component and $N(o; \mu, \Sigma)$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ, that is

$$N(o; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{-\frac{1}{2}(o-\mu)'\Sigma^{-1}(o-\mu)} \tag{8}$$

where n is the dimensionality of o. The exponent γ_s is a stream weight. It can be used to give a particular stream more emphasis.

3.2 ANN/HMM

There are two basic approaches to speech classification using neural networks: static and dynamic. In static classification, the neural network sees all of the input speech at once, and makes a single decision. By contrast, in dynamic classification which we used in this paper, the neural network sees only a small window of the speech, and this window slides over the input speech while the network makes a series of local decisions, which have to be integrated into a global decision at a later time [AT1989]. The neural network computes the weighted sum of its input and the passed this sum to a nonlinear function, most commonly a threshold or sigmoid function [RP1987]. The advantages of using neural network in HMM are the ability for discriminative training, no strong assumptions about the statistical distribution of the acoustic space, beter robustness to insufficient training data and ability to model acoustic correlation. It has been applied successfully to perform static pattern recognition or speech recognition [TK2009]. The neural network in our system is used as state emission probability estimator for HMM from a posterior probabilities.

A multilayer perceptron with fully connected neural network was used to model an output class which corresponding to phonetic unit. These posterior probabilities were used directly as the HMM state emission probabilities to perform a standard viterbi decoding which obtained the best phone sequence. Figure 1 shows the framework of acoustic model using MFCC and tone features as an input vectors. A feed-forward multilayer perceptron neural network we used is shown in Figure 2. There are 42 feature vectors per frame for the input layer with one hidden layer and all neurons use the non-linear sigmoid activation functions. The output of the neural network is a vector of posterior probabilities, with one class for each phone which is generated from input feature vectors.

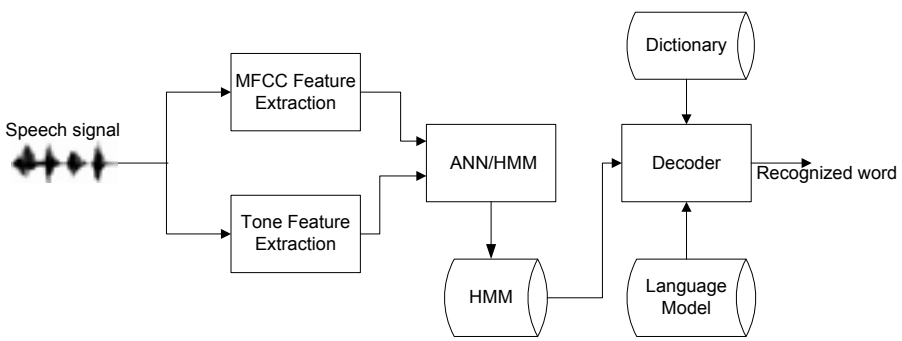

Figure 1. ANN/HMM acoustic training for HMM

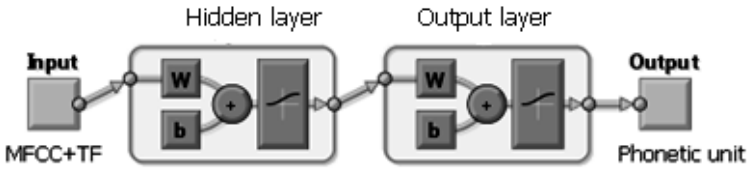

Figure 2. Neural network multilayer perceptrons

4 Experiment and Results

In this paper, the experiment was conducted using a speech corpus of ten Aesop's stories (translated into Thai) recorded from adult speakers. These stories contain all five tone levels in Thai language. The total number of words is 809 and the total number of distinct syllables is 2787. The data was collected from 10 native Thai speakers (5 male speakers and 5 female speakers), with different ages from 24 to 35 years old. The speech signals were sampled at 22 kHz and digitized with a 16 bit A/D converter. All speech data was recorded using Audacity program and stored as one sentence per file. We randomly split 80 percent of data to be used for training the model and 20 percent to be used for testing. Table 1 and Table 2 show the statistics of syllables and tone levels from the corpus. A GMM model was trained by using the Hidden Markov Toolkit [YS2002]. The parameter estimation was done by a flat start embedded training which requires the phonetic transcriptions of the available training utterances, while neural network feed-forward multilayer perceptrons was trained by Matlab. Both acoustic MFCC and Tone feature vectors are served as input vectors. The neural networks was trained on the same training data as the GMM/HMM systems with different input vectors and language model.

Story #	Number of syllable	Number of unique syllable
1	230	165
2	286	153
3	296	142
4	230	119
5	285	153
6	289	162
7	319	144
8	292	163
9	301	128
10	256	121

Table 1. Number of syllables of Aesop's stories

Tone	Number of syllable
Mid	687
Low	598
Falling	544
High	583
Rising	372

Table 2. Number of syllables for each tone of Aesop's stories

Configuration	WER(%)
MFCC + GMM + 2-gram	41.46
MFCC + GMM + 3-gram	31.15
MFCC + TF+GMM + 2-gram	26.35
MFCC + TF+GMM + 3-gram	24.76
MFCC + TF + ANN + 2-gram	25.20
MFCC + TF + ANN + 3-gram	23.35

Table 3. Experiment Results in term of Word Error Rate(%)

The results are summarized in table 3. The word error rates (WER) were reduced from the baseline HTK for both 2-grams and 3-grams of the language model. For the configuration of the combination of MFCC and tone features input, word error rates were reduced when ANN was applied to estimate the state emission probabilities in acoustic model. As the results, there are two difference things: Firstly, the different input features of acoustic model which we used pure MFCC and the combination of MFCC and tone feature vectors. Secondly, the method that uses to estimate the state emission probabilities. In this experiment, GMM and ANN were applied to estimate the state emission probabilities. Normally, speech recognition will gives some error especially in continuous speech because of the difficulty to segment the speech signal and the speaking speed. It shows more than 30 percent of word error rate for our corpus and around 20-25 percent for isolated word. When tone features were applied to be an input features, the recognition performance was improved more than 6% as shown in Table 3. Also the performance improved when ANN was applied with tone features. According to training data that we used in the experiments is not big enough, then the error might be occurred in most past of the adjacent syllable event there were difference tone. However, the language model is considered to greatly increase the performance of the continuous speech system as used in this experiment. Although, the different WER between ANN and GMM is not big enough, at least it showed that the proposed system can be improved the performance of speech recognition by reducing the word error rate.

5 Conclusions

In this paper, an approach based on the Artificial Neural Network (ANN) multilayer perceptrons is proposed to score the state emission probabilities under the HMM framework. A combination of the Mel-Frequency Cepstral Coefficients (MFCC) and tone information is used as input feature vectors to train an acoustic model. The total of 42 input vectors were normailzed and classified by multilayer perceptron neural network with 62 target outputs, each represents the phonetic units. The experiments were carried out to compare the performance between the ANN approach and the Gaussian Mixture Model (GMM) used by HTK with different language models. The results showed that the ANN approach with MFCC with tone features yielded a higher accuracy, i.e., lower word error rate (WER), for speech recognition compared to the GMM approach.

References

[AT1989] A. Waibel, T. Hanazawa, G. Hinton, K. Shiano, and K.Lang, Phoneme recognition using time-delay neural networks, In IEEE Trans. on Acoust., Speech, and Signal Processing, volume 37(3), pp. 328-339,1989.

[CT2006] Chutima Pisarn and Thanarak Theeramunkong, Improving Thai Spelling Recognition with Tone features, Springer-Verlag Berlin Heidelberg pp.388-398, 2006.

[DA1994] David M. Lubensky, Ayman O. Asadi, and Jayant M. Naik. Connected digit recognition using connectionist probability estimators and mixture-gaussian densities. In ICSLP, pp. 295--298.

[JR1997] John-Paul Hosom and Ronald A.Cole, A diphone-based digit recogition system using neural network, ICASSP-97,1997.

[MH1974] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, “Average magnitude difference function pitch extractor,” IEEE Transactions on Acoustics, Speech, Signal Processing, vol. ASSP22, pp. 353–362, 1974.

[NB2002] N. Thubthong and B. Kijsirikul, An empirical study for constructing Thai tone models, in Proc. the 5th Symposium on Natural Language Processing and Oriental COCOSDA Workshop, pp. 179–186, 2002.

[PA2008] Poonam Bansal, Anuj Kant, Sumit Kumar, Akash Sharda, Shitij Gupta. Improved model of HMM/GMM for speech recognition, "Intelligent Information and Engineering Systems" INFOS, pp.69-74, 2008.

[SM1999] S. Potisuk, M. P. Harper, and J. Gandour, Classification of Thai tone sequences in syllable-segmented speech using the analysis-by synthesis method, IEEE Transactions on Speech Audio Processing, vol. 7, no.1, pp. 95–102, 1999.

[SY1995] S.-H. Chen and Y.-R.Wang, Tone recognition of continuous Madarin speech based on neural networks, IEEE Transactions on Speech audio Processing, vol.3, no. 2, pp. 146–150, 1995.

[TC2003] T. Jeff Reynolds, Christos A. Antoniou, Experiments in speech recognition using a modular MLP architecture for acoustic modelling, Inf. Sci. 156 pp. 39-54, 2003.

[TK2009] Tom´aˇs Pavelka and Kamil Ekˇstein , A Comparison of Acoustic Models Based on Neural Networks and Gaussian Mixtures, Springer-Berlin / Heidelberg, pp.291-298, Volume 5729/2009.

[XM2006] X. Lei, M. Siu, M. Ostendorf, and T. Lee, Improved Tone Modeling for Mandarin Broadcast News Speech Recognition. Interspeech, 2006.

[YM1997] Younghong Yan, Mark Fanty and Ron Cole, Speech Recognition using neural networks with forward-backward probability generated targets, ICASSP-97,Vol.4 1997.

[YS2002] Young, S., et al.: The HTK Book, Cambridge University Engineering Dept, 2002.