Thai Word Recognition Using Hybrid MLP-HMM

Maleerat Sodanil[†] **Supot Nitsuwat**[††] **Choochart Haruechaiyasak**[†††]

[†], [††]Department of Information Technology King Mongkut's University of Technology North Bangkok, Thailand.
[†††]Human Language Technology Laboratory, National Electronics and Computer Technology Center (NECTEC), Ministry of Science and Technology, 112 Science Park, Klong Luang, Pathumthani 12120 Thailand.

Summary
The Hidden Markov Model (HMM) is a popular model for speech recognition systems. However, one of the difficulties in applying HMM is the estimation of the emission probabilities for constructing the Gaussian Mixture Models (GMMs). In this paper, we propose a method to estimate the state emission probabilities in HMM framework using Artificial Neural Networks (ANNs), particularly the Multi-Layer Perceptrons (MLPs). The proposed method can be considered as a hybrid MLP-HMM. Furthermore, tone information is one of highly potential features which could increase the recognition accuracy of tonal languages such as the Thai. Therefore, both MFCC features and tone features were extracted and served as the inputs for the MLP-HMM and the tone classifier. The posterior probabilities of outputs for each phone are represented as the state emission probabilities of the continuous density HMM framework. The experimental results showed that using the proposed hybrid MLP-HMM to train a Thai word recognition model helped improve the performance over the baseline system in terms of word error rates.

Key words:
Hybrid MLP-HMM, Thai speech recognition, tonal language.

1. Introduction

Some previous works in speech recognition system for tonal languages were proposed in order to improve the performance of speech recognition using additional information such as tone features. Previous experiment results showed that using tone features as additional inputs for training the acoustic model yielded higher accuracy compared to the baseline system for Thai [1] and Mandarin news broadcast speech recognition [2]. The method of context-independent acoustic model for Thai language has also been investigated [3]. The method of creating an acoustic model is considered to enhance the performance of learning from speech data. Hidden Markov Model (HMM) is well-known and popular in acoustic training data. The parameters of the model can be estimated and adapted automatically to give optimal performance. Although, HMMs are effective approaches to the problems of acoustic modeling, they also suffer from some limitations, for example, HMMs assumes the duration of exponential distribution, the transition probability depends only on the origin and destination, and all observation frames are dependent only on the state that generated them, not on neighboring observation frames. Furthermore, Gaussian Mixture Models (GMMs) are powerful when generating statistic values in the HMM frameworks. Neural networks have been used also in speech recognition with forward-backward probability generated targets [4],[5]. However, the connectionist-HMM framework which uses neural networks to generate the output posterior probabilities, which can be used to replace the GMMs acoustic model with a neural network to estimate the posterior probabilities of phonetic unit given the input vector of context window frames [6],[7]. It can be applied for continuous speech recognition [8] or integration with fuzzy logic in Arabic speech recognition [9]. To determine the model, first order left to right HMM models with self loops are generally used for acoustic models. An efficiency model for speech utterance is the Continuous Density Hidden Markov Model (CDHMM) which is suitable for describing the speech events [10]. In this paper, the state emission probabilities are estimated with an Artificial Neural Network particularly Multi-Layer Perceptrons (MLPs) so called Hybrid MLP-HMM in order to improve the performance of speech recognition over the HMM framework. The state emission probabilities of phoneme HMM will be estimated from the output node of the MLP. Then Viterbi algorithm is employed to be used as the decoder. Tone features are extracted from speech signal and classified by MLP as additional feature for tonal languages. The comparison of the baseline system is tested with different configurations, such as tone features and a number of hidden layers in the MLP classifier throughout the experiments.

The rest of this paper is organized as follows. In Section 2, a review of the Thai phonetic system is presented. In Section 3, the proposed framework consisting of a hybrid MLP-HMM and tone recognition will be introduced. The experiment and results are described in Section 4. Section 5 gives the conclusion.

Manuscript received March 5, 2010
Manuscript revised March 20, 2010

2. Review of Thai Phonetic System

2.1 Thai Syllable Structure

The Thai language is a tonal language which makes it very different to Western languages. The phonetic structure of Thai is based primarily upon the monosyllable. Thai syllable structure consists of $/C(C)V(:)(C)^T/$ where C,V,: and ^{T} represent an initial consonant (cluster consonant), vowel (short or long) and lexical tone, respectively [11]. Each syllable perceived a choice between five distinct tones which separates into two groups. Firstly, level tone which consists of Mid, Low and High tone. Secondly, contour tone which consists of Falling and Rising. Fig 1 and Table 1 shows characteristics of five Thai tones and the examples of each tone.

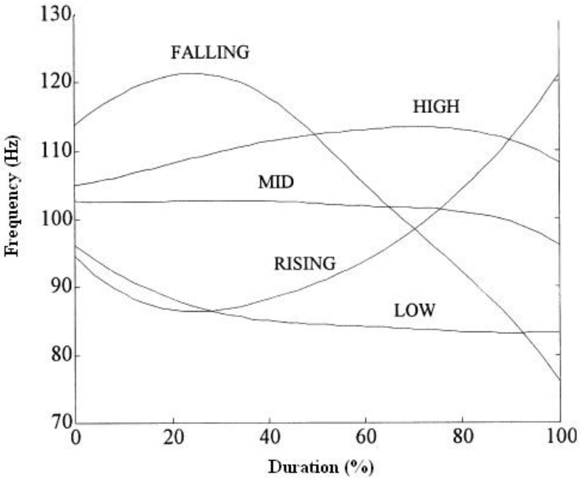

Fig. 1 F_0 contour of five tones in standard Thai language

Table 1: Example of word for each tone

Tone Group	Tone	Word (IPA)	Word in Thai
Level	Mid	/jaːw^0/	ยาว
	Low	/buak^1/	บวก
	High	/ruː^3/	รู้
Contour	Falling	/baːn^2/	บ้าน
	Rising	/suaj^4/	สวย

2.2 Factors determining the tone

There are two different types of Thai syllable: stressed and unstressed syllable. Stressed syllable can be stressed individually in normal speaking. Its consists of at least three parts: initial consonant vowel and tone as (CV^T) or a maximum of five parts, two initial consonants (cluster consonant), one vowel, one final consonant and tone ($CCVC^T$). The contrast of unstressed syllable appears in the unstressed position of normal speaking. The Thai phonemes set as shown in Table 1 consists of 21 initial consonants, 11 cluster consonants, 9 final consonants, 21vowels (short vowel, long vowel and diphthong) and 5 tones(4 symbols) which is represented by the number 0,1,2,3 and 4 respectively. The 44 Thai consonants are divided up into three groups known respectively as High, Middle and Low class consonants as shown in Table 2. There are many cases where the letter ห (h) in a sample word หมอน (mɔn^4) as an initial consonant is silent and there are a few cases where the letter อ (?) in a sample word of อย่า (jaːk) as an initial consonant is also silent, but this makes no difference to the rule, the tone is still governed by the class of the initial consonant even though it can be a silent consonant.

Table 2: Thai Phoneme Set (IPA)

Phoneme Type	Phoneme Set
Initial Consonant (21)	p t c k ? ph th ch kh b d m n ɳ l r f s h w j ป ต จ ก อ พ ท ช ค บ ด ม น ง ล ร ฟ ซ ฮ ว ย
Cluster Consonant (11)	pr phr pl phl tr kr khr kl khl kw khw ปร พร ปล พล ตร กร คร กล คล กว คว
Final Consonant (9)	-p -t -k -? -m -n -ɳ -j –w บ ต ก อ ม น ง ย ว
Short vowel (9)	i e ɛ ɯ ɤ a u o ɔ ◌ิ เ-ะ แ-ะ ◌ึ เ-อะ -ะ ◌ุ โ-ะ เ-าะ
Long vowel (9)	iː eː ɛː ɯː ɤː aː uː oː ɔː ◌ี เ- แ- ◌ื เ-อ —า ◌ู โ- —อ
Diphthong (3)	ia ɯa ua เ◌ีย เ◌ือ ◌ัว
Tone (5)	0 1 2 3 4 (- ◌่ ◌้ ◌๊ ◌๋)

Table 3: Initial consonants class and consonant members

Class	Phoneme (Consonant - Phone represented)
High	ข-kh ฉ-ch ฐ ถ-th ผ-ph ฝ-f ศ-ษ-ส-s ห-h
Middle	ก-k จ-c ฎ-ด-d ฏ-ต-t บ-b ป-p อ-z
Low	พ-ภ-ph ฟ-f ฑ-ฒ-ท-ธ- th ค-ฅ-ฆ-kh ซ-s ฮ-h ช-ฌ-ch ง-ng ญ-ย-j น-ณ-n ร-r ว-w ม-m ฬ-ล-l

Table 3 shows all consonants with 21 represented phones. In the pronunciation, we used a phone to represent the sound. Therefore, some words can have the same sound of phone but different transcription. All words which do not end in a vowel sound must have either -p -t -k -? -m -n -ɳ -j –w, which is represented บ ต ก อ ม น ง ย ว as the final sound. Although this is strictly true, also in some conversation the final consonant is often slurred and particularly after a long vowel, the final "p" may sound more like a "b" and the final "t" sounds more like a "d".

3. The Proposed Framework

The proposed framework of Thai word recognition using hybrid MLP-HMM architecture is shown in Fig 2. There are 3 main parts: 1) Phone classification: MFCC feature vectors will be extracted from speech signals and fed into MLP, the output posterior probability for each nodes corresponding to the phonetic units, 2) Tone classification: tone features will be extracted from voiced portions and fed into MLP to classify five tones as an output, and 3) The combination module: combined those parts for the final output. The details for each part will be described later on.

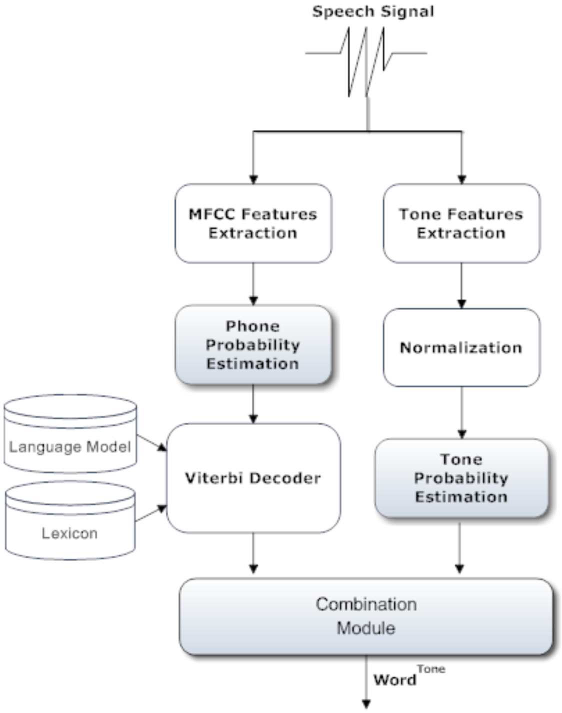

Fig. 2 Overview of a framework structure

3.1 Phone Classification

The Hidden Markov Model (HMM) is a powerful statistical method of characterizing the observed data samples of discrete-time series. It has been successfully used in automatic speech recognition. HMM is basically a markov chain where the output observation is a random variable X generated according to a output probabilistic function associated with each state. The definition of HMM is defined by $\lambda = (A, B, \pi)$ which A is a transition probability matrix, B is an output probability matrix and π is an initial state distribution. There are two categories of HMM based on statistical models: 1) Discrete HMM (DHMM) which evaluates probabilities based on discrete data counting and 2) Continuous Density HMM (CDHMM) which evaluates probabilities based on continuous Probability Density Functions (PDFs) that is usually referred to as likelihoods. DHMM uses a vector quantization based method for computing the state probability. For instance, frame i has to be converted into the corresponding symbol k = O(i), and the probability of symbol k to state j is retrieved from B(k, j) of the matrix B. CDHMM uses a continuous probability density function for computing the state probability. The method for identifying the optimum parameter that maximize the probability (likelihood) of the sample data is based on re-estimation of Maximum Likelihood Estimate (MLE). Although the computation of probabilities with discrete models is faster than with continuous models, CDHMM will be considered in order to solve the problem of discrete HMM during vector quantization process.

As the baseline system in general, Gaussian Mixture Models (GMMs) are the most popular and effective choice of PDFs for CDHMM model. In HTK [12], the output distributions represented by Gaussian Mixture Densities which 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

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

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

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

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.

Hybrid MLP-HMM Probability Estimation

Artificial Neural Network particularly Multi-Layer Perceptrons (MLPs) estimate the HMM state posterior probabilities was proposed by Bourlard et al, [13] which rely on a probabilistic interpretation of the MLPs outputs. Each output unit of MLP is trained to perform a non-parametric estimate of posterior probability of a left-to-right CDHMM state given the acoustic observations. This represents a fundamental class of hybrid models. MLPs are used to estimate the state emission probabilities required in HMM as shown in Fig 3. The MLPs in this work consists of M input and N output nodes separated by a number of hidden nodes. The output of each layer forms the input of the consecutive layer. Each output node represents one symbol class corresponding to the phonetic units. The discriminative training of the MLP is performed by the feed-forward algorithm. The initialization of the MLPs weights is chosen randomly.

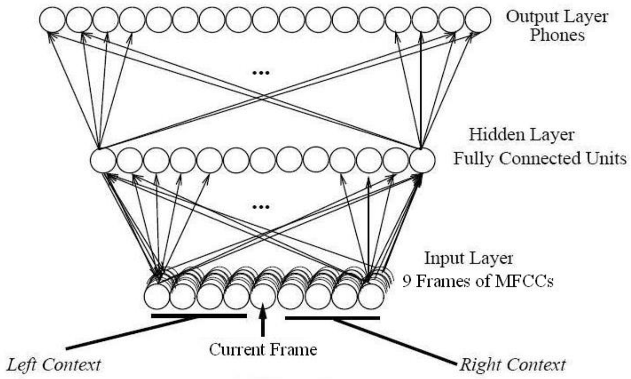

Fig. 3 Hybrid MLP-HMM framework

In this MLPs based phoneme classifier, 39 dimensional of MFCC feature vectors are served as an input data of MLP input layer which is fully-connected to one hidden layer. The output layer has 53 neurons corresponding to the Thai phonetic units. In the hidden layer, the sigmoid function is used as an activation function. A number of hidden nodes are changed through the experiments for the best results.

As all HMMs use the same MLP as source for their PDFs in continuous HMM, The output probability b_j of each state s_j of the HMM is computed by a weighted sum of a fixed number I of PDFs:

$$b_j(o_t) = P(o_t|s_j) = \sum_{i=1}^{I} c_{ji} \cdot P(o_t|\rho_i) \tag{3}$$

The conditional probability $P(o_t|\rho_i)$ of Eq. 3 is not available from the MLP. However, the output of the MLP resembles the posterior probability $P(o_t|\rho_i)$ To replace $P(o_t|\rho_i\ )$ by $P(\rho_i|o_t)$ we can use a scaled likelihood (the likelihood divided by the probability of the observation) [13]. This probability can be expressed with the posterior probabilities $P(\rho_i|o_t)$ and the priori class probabilities $P(\rho_i)$, which can be estimated using the training data. The scale likelihoods are estimated by applying Bayes' Rule to the MLP output as:

$$P(\rho_i|o_t) = \frac{P(o_t|\rho_i)P(\rho_i)}{P(o_t)} \tag{4}$$

then

$$\frac{P(o_t|\rho_i)}{P(o_t)} = \frac{P(\rho_i|o_t)}{P(\rho_i)} \tag{5}$$

Eq. 4 and 5 lead to the tied posterior approach in which the output probabilities b_j can be computed as:

$$b_j = P(o_t|s_j) = \sum_{i=1}^{I} c_{ji} \cdot \frac{P(\rho_i|o_t)}{P(\rho_i)} \tag{6}$$

The output of the classifier estimates the posterior probabilities of the target classes given the input. Therefore the HMMs are kept by the PDFs used in Eq. 6 generated by the MLP. In this work, we use phoneme as classes, posterior features are estimates of phonemes posterior probabilities given the MFCC based features.

The posterior-based speech feature is formed by output values at each time frame. The net input consists of 9 consecutive frames as the contextual aspects which are known to play an important role in speech recognition [14] will be taken into account. At the time frame t, a left and right context is typically four frames, (o_{t-4}, …, o_t, …, o_{t+4}), are used as input for the MLP. Then, the output values represent the posterior probability of the phoneme ρ_i given the input, i.e. each output is an estimate of $P(\rho_i|\ o_{t-4}, \ldots, o_t, \ldots, o_{t+4})$. The posterior feature vector is then formed by the set of MLP output values $\{P(\rho_1|o_t), \ldots, P(\rho_i|o_t)\}$ where i is represent the total number of phonemes which corresponds to the number of output nodes. The training algorithm procedure is:

- CDHMM training using Viterbi decoding
- MLPs training as phoneme classifier using feed-forward networks
- Re-training the MLPs with the iterative scheme in which the model obtains the data from 2 steps above, and used to classify the training data. The discriminative criterion function aims for high performance, typically based on mean squared error criterion (MSE) which measures the average square difference between the model output and the desired output. The Mean Squared Error (MSE) criterion is defined as:

$$E = \sum_{n=1}^{N} ||g(x_n) - d(x_n)||^2 \tag{7}$$

where x_n is the pattern to be classified, $d(x_n) = (d_1(x_n), \ldots, d_k(x_n), \ldots, d_K(x_n))^t$ represents the desire output vector (for classes $\rho_k, k = 1, \ldots, K$), $g(x_n) = (g_1(x_n), \ldots, g_k(x_n), \ldots, g_K(x_n))^t$ the observed output vector, K the total number of classes, and N the total number of training patterns.

Three layered MLPs have been used in our experiments. The activation function of the hidden layer is the logistic function which is one of the most common sigmoid functions as:

$$f(x) = \frac{1}{1 + \exp(-ax)} \qquad (8)$$

where a is a constant controlling the slope of the function. The output activation function is the softmax function as:

$$f(x_i) = \frac{\exp(x_i)}{\sum_{n=1}^{K} \exp(x_i)} \qquad (9)$$

where K is the number of categories (phoneme units) in the output layer. In our framework, the softmax activation function is implemented using exponential units, multiplication units and an inverter unit, to ensure that the output activities sum to one.

The main advantages of using a Hybrid MLP to estimate the state emission distributions are discriminative among the output classes which the scaled likelihoods are estimated to be maximized for the right class, the capability of the hidden layer to model high order moments helps in modeling correlation within an acoustic feature vector and across acoustic feature vectors over time when feeding the MLP with a consecutive contextual of speech features, and a better model with irregular class boundaries on the acoustic space as a non-linear classifier.

3.2 Tone Classification

The fundamental frequency (F_0) or pitch can be extracted from only the voiced part of the time unit in the utterance. 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 [2]. In this paper, the average magnitude difference function (AMDF) is used instead of autocorrelation function to extract the pitch period, as used in [15]. It computes the difference between the signal and time shifted version of itself. The average magnitude difference function [16] is defined as:

$$AMDF(\tau) = \frac{1}{N} \sum_{n=0}^{N-1-\tau} |x(n) - x(n-\tau)| \qquad (10)$$

where *x(n)* are the samples of analyzed speech frame. *x(n+τ)* are the samples time shifted τ seconds and N is the frame size.

Since F_0 is present only in voiced segments, it needs to be interpolated in unvoiced regions in order to avoid variance problems in recognition. In this paper, the moving average smoothing algorithm has been applied. The moving average smoothing is defined as follows:

$$\hat{F_n} = \frac{1}{N} \sum_{i=n-N/2}^{n+N/2} F_i \qquad (11)$$

where F_i is the order I of F_0, $\hat{F_n}$ 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

$$delta_n = \begin{cases} \dfrac{f_{n+\theta} - f_{n-\theta}}{2\theta}, & \theta < n < N - \theta \\ f_{n+1} - f_n, & n < \theta \\ f_n - f_{n-1}, & n \geq N - \theta \end{cases} \qquad (12)$$

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 the total frame. The total of tone feature equal 3 feature vectors for each frame.

To classify five standard Thai tones, all tone features ($\hat{F_n} + \Delta\hat{F_n} + \Delta\Delta\hat{F_n}$) are then normalized to lie between -1 and +1 to be used in MLP tone classifier using the following equation:

$$norm\ F_i = 2.0 * \left(\frac{F_i - minF_i}{maxF_i - minF_i}\right) - 1.0 \qquad (13)$$

where F_i is the *i*th feature under consideration, $minF_i$ and $maxF_i$ are the minimum and maximum value of F_i

MLP tone classifier consists of 3 feature vectors of duration points between 0 and 100 percent step of 10 percent of voiced portion. One hidden layer with fully-connected was presented and 5 output target corresponding to Thai tone levels as shown in Fig 4.

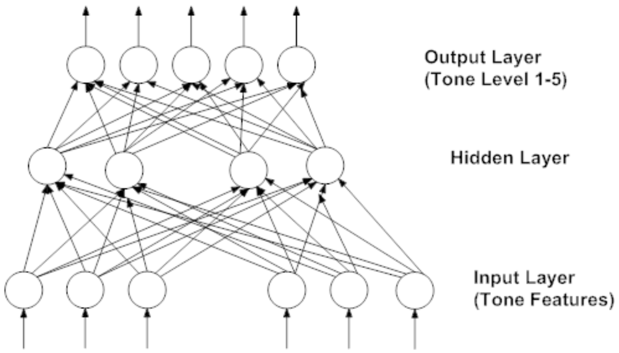

Fig. 4 MLP tone classifier

3.3 Combination Module

The combination module works as word summarized using the output of phoneme recognition and tone recognition module as an input data. The output will be verified by the monosyllable pronunciation database which is generated from syllable rule based defined as /C(C)V(:)(C)^T/ where C,V,: and ^{T} represent an initial consonant (cluster consonant), vowel (short or long) and lexical tone respectively.

4. Experiment and Results

4.1 Speech Database

A group of sample words for each initial consonant, vowel and final consonant were recorded and transcribed. The data set was collected from 10 native Thai speakers (5 males and 5 females) in which each speaker pronounces the same word 10 times. All speech signals were sampled at 22 kHz digitized with a 16 bit A/D converter using Audacity program[1].

4.2 Experimental Setup

Feature extraction for phone based recognizer, 39 dimensional MFCC feature vectors (12 MFCC plus energy and their first and second order temporal derivatives) were extracted from speech signals with pre-emphasis performed first. Speeches are analyzed based on a frame size of 25ms and shifted window of 10ms using hamming window. The baseline system of automatic speech recognition was compared, a continuous phone-based HMM recognizer was implemented using HTK [11] for comparison purposes. Each phone was represented as a 5 state left-to-right model with one Gaussian mixture using diagonal co-variances. The acoustic models were trained using maximum likelihood estimator (MLE) as a statistical method to estimate the value of parameters, based on a set of observations of a random variable that related to the parameters being estimated..

Phoneme recognition using hybrid MLP-HMM framework, 39 dimensional MFCC feature vectors are fed into input layer with fully-connected to one hidden layer and 53 output neurons corresponding to phonetic units. In order to provide the MLP with contextual information, 9 consecutive frames of data are given as input.

Tone recognition, the feature set of input layer is based on a smoothed F_0 and delta, double delta F_0 with fully-connected to one hidden layer and 5 output neurons corresponding to five tones.
Additional, bigram model is used as language models for all configurations in part of decoder.

4.3 Performance Measure

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:

$$WER = \frac{N - S - D - I}{N} \times 100\% \qquad (14)$$

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

4.4 Results

The performance of hybrid system was compared with an HMM-based recognizer where emission probabilities are modeled with mixtures of Gaussian components. During training using MLPs, the numbers of hidden nodes were varied to find the best accuracy as shown in Table 4. Table 5 shows the accuracy for each tone from the output of tone classifier. Table 6 shows the results of word recognition in terms of word error rate.

Table 4: The accuracy of phone classifier by varying the number of hidden nodes (%)

No. Hidden Node	Performance	Accuracy
500	MSE	84
500	SSE	76
1500	MSE	72
2000	MSE	68

Table 5: The accuracy results of tone recognition

Tone	Accuracy (%)
Mid	97.45
Low	96.56
High	93.32
Falling	95.50
Rising	97.35

Table 6: The accuracy results in terms of Word Error Rate (%)

Configuration	WER (%)
Baseline	25.6
MLP-HMM	21.2
Tone+Baseline	20.3
Tone+MLP-HMM	19.5

[1] http://audacity.sourceforge.net/

The experiments showed that training using MLP posterior probabilities as state emission probabilities in HMM proposed as hybrid architecture improved recognition performance over the baseline CDHMM.

5. Conclusion

In this paper, we proposed a hybrid MLP-HMM approach with tone recognition in order to improve the performance of Thai automatic speech recognition. The proposed approach consists of two main components: (1) a hybrid MLP-HMM as part of an acoustic model and (2) a tone feature extraction and classification using MLPs. The emission probabilities in the HMM framework are estimated by the posterior probabilities of neural network multilayer perceptrons in which the MFCC feature vectors are served into an input layer with fully-connected hidden layer. All nodes of output layer are represented by phonetic units. The MLPs are trained to estimate the emission probabilities for each state probability of the continuous density HMM framework. The Viterbi decoder is then used to find the single best state sequence using a dynamic programming algorithm. Compared to the baseline GMM system, the hybrid MLP-HMM significantly improved the performance of context-independent networks. The recognition rates for consonants and vowels from the hybrid MLP-HMM model are higher than the baseline model, because the MLP model could help generate a recognition model which is more discriminative than the baseline HMM. In summary, the overall recognition rate was improved by using the proposed hybrid MLP-HMM approach.

References

[1] C. Pisarn and T. Theeramunkong, “Improving Thai Spelling Recognition with Tone Features”, Springer-Verlag Berlin Heidelberg, pp. 388-398, 2006.

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

[3] S. Kasuriya, S. kanokphara, N. Thatphithakkul, P. Cotsomrong and T. Sunpethniyon, “Context-independent Acoustic Models for Thai Speech Recognition”, ISCIT2004, pp. 991-994, 2004.

[4] Y. Yan, M. Fanty and R. Cole, “Speech Recognition using Neural Networks with Forward-Backward Probability Generated Targets”, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97), vol 4, 1997.

[5] P.Van Tuan and G. Kubin, “DTW-Based Phoneic Groups Classification using Neural Networks”, ICASSP, pp. 401-404, 2005.

[6] S. Renals and N. Morgan, “Connectionist Probability Estimation in HMM Speech Recogniton”, International Computer Science Institute, 1992.

[7] P. D. Polur and G. E. Miller “Investigation of an HMM/ANN hybrid structure in pattern recognition application using ceptral analysis of dysarthric (distorted) speech signals”, pp. 741-748, 2006.

[8] E. Trentin and M. Gori. A survey of hybrid ANN/HMM models for automatic speech recognition. *Neurocomputing*, 37(1/4); 91-126, March 2001.

[9] C.Octavian DUMITRU and I. GAVAT, “Vowel, Digit and Continuous Speech Recogniton Based on Statistical, Neural and Hybrid Modelling by using ASRS_RL”, The International Conference on Computer as a tool, pp. 856-863, 2007.

[10] P. Perner and A.Rosenfeld, “Connectionist Probability Estimators in HMM Arabic Speech Recognition Using Fuzzy Logic”, Springer-Verlag Berlin Heidelberg, pp.379-388, 2003.

[11] S.Tangwongsan, P. Po-Aramsri and R. Phoophuangpairoj, “Highly Efficient and Effective Techniques for Thai Syllable Speech Recognition”, Springer-Verlag Berlin Heidelberg, pp. 259-270, 2004.

[12] K. Naksakul, “Thai Phonology System”, Chulalongkorn University, revised 2008.

[13] Young, S., et al., “The HTK Book”, Cambridge University Engineering Dept, 2002.

[14] H. BOURLARD, N.MORGAN “Connectionist speech recognition, Kluiwert”, Academic Publishers, 1994.

[15] S. Furui, “Speaker independent isolated word recognizer using dynamic features of speech spectrum”, IEEE Trans. on Acoustic, Speech, and Signal Processiong, vol. 34, no. 1, pp. 52-59,1986.

[16] Thubthong, N “A method for isolated Thai tone recognition using a combination of neural networks”. Computational Intelligence, Volume 18, Number 3, 2002, pp. 312-335.

[17] 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.

Maleerat Sodanil received B.Sc. degree in Computer Education from King Mongkut's University of Technology North Bangkok Thailand, M.S. degree in Computer and Information Technology from King Mongkut's University of Technology Thonburi Thailand. She is a Ph.D. candidate in the Department of Information Technology at King Mongkut's University of Technology North Bangkok. Her current research interests Speech Recognition System, Data Mining.

Supot Nitsuwat received B.S. in Mathematical Physics from Ramkhamhaeng University Thailand, M.S. in Applied Mathematics from Mahidol University Thailand and Ph.D. degree in Computer Science from The University of New South Wales, Sydney, Australia. His current research interests Multimedia Information Retrieval, Pattern Recognition, **Knowledge** Representation and Reasoning. Currently, he is an executive management in the Department of Information Technology at King Mongkut's University of Technology North Bangkok, Thailand.

Choochart Haruechaiyasak received B.S. from University of Rochester, M.S. from University of Southern California and Ph.D. degree in Computer Engineering from University of Miami. His current research interests Search technology, Data/text/Web mining, Information filtering and Recommender system. Currently, he is chief of the Intelligent Information Infrastructure Section under the Human Language Technology Laboratory (HLT) at National Electronics and Computer Technology Center (NECTEC), Thailand.