Tone Recognition Based on Multi-layer Perceptron with Application for Continuous Mandarin Speech Recognition

Wern-Jun Wang[1], Chih-Heng Lin[1] Chien-Hung Chen[1,2], I-Bin Liao[1,2] and Eng-Fong Huang[1]

[1]*Multimedia Application Technology Laboratory, Chunghwa Telecom. Labs, Taiwan, R.O.C.*
[2]*Department of Communication, National Chiao Tung University, Taiwan, R.O.C*
{wernjun, chih, cchdavis, snet, engfong}@cht.com.tw

ABSTRACT

Some multi-layer perceptron (MLP) based tone recognition schemes for continuous Mandarin speech are discussed in this paper. The basic MLP scheme using both local and contextual features is first employed for tone recognition. The modular MLP scheme is then introduced to increase the discriminability of the basic MLP tone recognizer. Moreover, we also investigate how to incorporate tone recognizer into large vocabulary continuous speech recognition. A two-pass, post-processing scheme is proposed to utilize the recognized tones in rescoring the recognized N-best strings. The effectiveness of these schemes was confirmed by simulations on a mobile ring back tone (RBT) service with automatic speech recognition function. It has been found from the experimental results that weighted tone information can yield 6% relative improvement of the recognition rate.

1: INTRODUCTION

Mandarin Chinese is a tonal and syllabic language. Each Chinese character is pronounced as a monosyllable with which a tone associates. All monosyllables have a very regular, hierarchical phonetic structure [1] which is composed of a base-syllable and a tone. A base-syllable can be further decomposed into optional syllable *initial* and syllable *final*. Basically, there are five lexical tones which are commonly labeled in sequence from Tone 1, Tone 2, Tone 3, Tone 4 to Tone 5. Notice that different characters may have the same base-syllable but with different tones. It implies that the tonality of a monosyllable is also lexically meaningful. Conventionally, a complete continuous Mandarin speech recognition system is generally composed of acoustic decoding for mono-syllable identification and linguistic decoding for word (or character) string recognition. Owing to the regular hierarchical phonetic structure of mono-syllables, acoustic decoding is traditional further decomposed into base-syllable recognition and tone recognition.

Many studies have been conducted on how to incorporate the tone information into continuous Chinese speech recognition. The approaches can be roughly divided into two major categories. One is to train tone dependent acoustic models by including the pitch-related features as extra components in the short-time acoustic feature vector [2][3]. The other is to implement tone recognition and acoustic decoding separately. The result of tone recognition is then incorporated with the acoustic decoding in a post-processing stage [4][5]. Similar to the latter approach, the study of tone recognition and its incorporation into continuous Mandarin speech recognition are discussed in this paper. Some multi-layer perceptron (MLP) schemes are proposed to build a tone recognizer. A two-pass, post-processing scheme is then adopted for using tone recognizer to reorder the N-best hypotheses generated by acoustic decoding and effectively improve the recognition rate.

2: MULTI-LAYER PERCEPTRON FOR TONE RECOGNITION

In the training phase of our proposed MLP-based tone recognizer, each training utterance is first processed in acoustic decoding to obtain the best syllable-boundary segmentation matching with the associated text. Some recognition features are then extracted based on the best syllable-boundary segmentation. A basic MLP tone model is then trained by the backward error propagation algorithm to learn the relationship between the input features of the training utterance and the tone information of the associated text. In the testing phase, the input utterance is first processed in acoustic decoding to generate top-N hypothesis strings. Some recognition features are then extracted based on the syllable-boundary segmentation of each hypothesis. The well-trained MLP tone recognizer is then employed to

generate the recognized tone results by using these input features.

2.1: THE FEATURES FOR TONE RECOGNITION

The features used for tone recognition are extracted from the vicinity of the processing syllable. The recognition features include two subsets. One contains some local features of the current syllable segment, while the other contains some contextual features extracted from the syllable segment and its two nearest neighbors. Fig. 1 shows a schematic diagram of the feature extraction. Local features in the first set include:

(L1) the duration of the pitch contour of the syllable (d_c),
(L2) the mean of three uniformly divided log-energy sub-contours (EM_{c1}, EM_{c2}, EM_{c3}),
(L3) the mean and slope of three uniformly divided pitch sub-contours (PM_{c1}, PS_{c1}, PM_{c2}, PS_{c2}, PM_{c3}, PS_{c3}),
(L4) the correlation coefficients of three uniformly divided sub-segments (CC_{c1}, CC_{c2}, CC_{c3}).

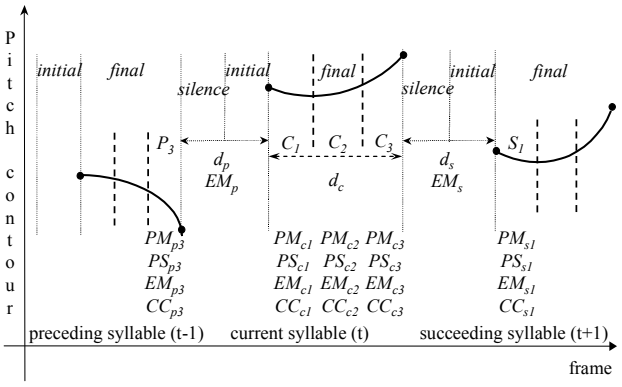

Figure 1: A schematic diagram showing the extraction of the input features for tone recognition.

The second set contains contextual features extracted from the current syllable segment and its two nearest neighbors. They include:

(C1) the four features, i.e. log-energy (EM_{p3}), pitch mean (PM_{p3}), pitch slope (PS_{p3}) and correlation coefficient (CC_{p3}), extracted from the last sub-segment of the preceding syllable,
(C2) the four features, i.e. log-energy (EM_{s3}), pitch mean (PM_{s3}), pitch slope (PS_{s3}) and correlation coefficient (CC_{s1}), extracted from the first sub-segment of the succeeding syllable,
(C3) log energies and durations of both unvoiced/silence segments before (EM_p, d_p) and after (EM_s, d_s) the current syllable,
(C4) two binary indicators used to indicate the boundary condition of the processing syllable.

Here the subscripts p and s stand for the preceding and the succeeding syllables, respectively. The reasons of using these features are discussed as follows. The eight features in (C1) and (C2) are the main features used to deal with coarticulation. The tightness of relationships between the current syllable and the two nearest neighbors are implicitly described by the four features in (C3). The last two features in (C4) are two Boolean flags. They are used to indicate whether the processing syllable is the first or last syllable of the input utterance.

After combining the two sets of recognition features, there are in total 27 features extracted for each syllable segment. In the feature-assembling process, two things need to be specially taken care. One is the setting of boundary conditions and the other is the normalization operation for all features before training the MLP. The detailed descriptions of theses two considerations are as follows.

The simple boundary-condition-setting scheme

All non-existing features for some syllable segments in the beginning and ending parts of every utterance are set to zero. Investigating the responses of the MLP tone model to the boundary conditions confirm the suitability of this scheme.

The normalization operation

The mean and standard deviation of the feature vector are computed based on the statistics of the training corpus. Each element of the normalized feature vector is then obtained by first subtracting its corresponding mean from its original value and then being divided by its corresponding standard deviation. From our pilot experiment, the MLP based tone recognizer trained by the normalized feature vectors outperformed the one trained by the un-normalized feature vectors.

2.2: BASIC MLP

The above-mentioned features are then fed into an MLP pattern recognizer for tone recognition. The basic MLP used in this paper is a two-layer network with a single hidden layer. It consists of four output nodes corresponding to four tones. Notice that the neutral tone, i.e. Tone 5, is often considered as the shortened counterparts of Tone 3 [1]. Besides, the tonalities of most content words don't generally include Tone 5, especially for the vocabulary used in specific domain. Therefore, in the following experiments, only four distinctive tone categories are defined. Moreover, each neuron output is the sigmoid function of the weighted summation of inputs. The MLP recognizer is then trained by the backpropagation algorithm, which

minimizes the mean squared error between the feedforward outputs and the desired targets.

2.3: MODULAR MLP

To increase the discriminability of traditional MLP, some modifications of MLP have been proposed for different applications. For example, the benefits of decomposing the MLP have been shown for speaker recognition task [6] and hand-written OCR problems [7]. The decomposition of MLP is conceptually straightforward. Suppose we have N classes, then instead of employing 1 MLP with N outputs, we use N MLPs, each with a single output. Such networks used for tone recognition can be trained with one 'genuine' tone data and several 'imposter' tone data. This scheme can be categorized to one class one network (OCON) model, in which the imposter tone data are used as counterexamples. An important issue of the OCON model training is to balance the number of the class specific data and the number of the counterexamples. Thus the degree of unbalancing can affect the performance of the OCON based pattern recognizer and usually result in performance degradation. To be free from the performance degradation problem of the OCON based recognizer, the modular MLP is therefore proposed [6]. The modular MLP is the combination of several expert networks and each expert network is trained with the genuine tone and one of the imposter tones.

The modular MLP structure used for tone recognition was shown in Fig. 2. It includes 12 MLP based expert networks. The symbol embedded in the MLP box, for example, "MLP {A, ~B}" means that the corresponding MLP is trained by using the training syllables with Tone A and Tone B. The desired output responses are set to 1 for the syllables with Tone A, while the desired output responses are set to 0 for the syllables with Tone B.

In Fig. 2, the decision logic is designed to combine the outputs of above-mentioned twelve well-trained MLP based expert networks. The decision logic can be implemented in many different ways. A simple voting method is chosen in our study. The detailed descriptions of the operations of modular MLP for tone recognition are as follows.

- Calculate the normalized feature for the processing syllable according to the normalization operation described in Section 2.1.
- The normalized features of the processing syllable are fed into every MLP based expert network.
- For a MLP with symbol "MLP {A, ~B}", we get one vote for Tone A if the output of the MLP is greater than a predefined threshold; otherwise, we get a vote for Tone B.

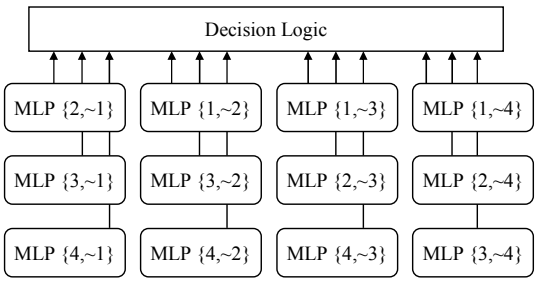

Figure 2: Modular MLP structure for tone recognition.

- Finally, the tonality with the maximum votes is decided for the processing syllable.

2.4: EXPERIMENTAL RESULTS OF TONE RECOGNITION

A large multi-speaker database was used for the study of tone recognition. It was recorded by 242 female and 243 male speakers. Each speaker read several words and short phrases with length ranging from 2 to 10 characters. All the utterances were collected over the GSM telephone network. The database was divided into two parts, one for training and one for testing. The training set contained about 84,000 syllables uttered by 217 female and 218 male speakers. The test set contained about 10,600 syllables uttered by other 25 female and 25 male speakers.

The experimental results were shown in Table 1. The performance of modular MLP is slightly better than that of basic MLP. Besides, the node number of hidden layer and the required training iteration number have been added for comparison in this table. It is clear that the advantages of modular MLP were shown on the fewer number of hidden nodes and convergence in fewer training iterations. In addition, another advantage is that the algorithm can distribute the forward feeding process over many machines by adopting parallel processing scheme. With respect to further improve the performance of modular MLP, two things still need to be carefully investigated. One is the training sample unbalancing problem and the other is the more sophisticated method to implement the decision logic.

Table 1: Tone recognition rate of the proposed MLP based methods

	Node number of hidden layer	Training iteration number	Accuracy
Basic MLP	50	20	81.5%
Modular MLP	20	12	82.4%

3: INTEGRATING TONE RECOGNITION INTO CONTINUOUS MANDARIN SPEECH RECOGNITION

The MLP-based tone recognition has been integrated into the mobile ring back tone (RBT) system developed at Chunghwa Telecommunication (CHT) Company. The mobile RBT system provides the customers to download their favorite tunes with speech recognition approach. To use this service, users can easily select their favorite tunes by speaking the song title or the singer name using cellular phone. It is one of the intelligent value-added services developed at Chunghwa Telecommunication Laboratories by applying state-of-the-art speech signal processing techniques to CHT fixed and mobile network business. The mobile RBT service was deployed in Taiwan since 2005.

3.1: BASELINE CONTINUOUS MANDARIN SPEECH RECOGNITION SYSTEM

The baseline continuous Mandarin speech recognition discussed in this paper is the above-mentioned mobile RBT system. The acoustic models are context-dependent syllable *initial* and context-independent syllable *final*. The acoustic feature vector is composed of 12 MFCC, 12 delta MFCC and one delta energy. The cepstrum mean subtraction (CMS) technique has been used to minimize the channel (transducer) mismatch and the speaker variability. The state numbers of *initial* and *final* model are, respectively, 3 and 5. The mixture numbers range from 4 to 32 depends on the number of training samples. The lexicon used in the RBT task includes two major sets. The first set includes singer name only with vocabulary size 1789. The second set includes both singer name and song title with vocabulary size 47089.

3.2: USE OF TONE RECOGNITION FOR POST-PROCESSING OF CONTINUOUS SPEECH RECOGNITION

According to the experimental results shown in Section 2.4, the performance obtained by modular MLP scheme is just slightly better than basic MLP scheme. Therefore, considering the tradeoff between the required computation time and the performance gain, only the basic MLP-based tone recognizer was used in the following experiment. A functional block diagram of our proposed two-stage scheme was shown in Fig. 3. The N-best hypothesis strings and their segment boundary information were generated after the multi-keyword spotting process. Each hypothesis may include one or two keywords, i.e. the title and the optional singer name of the desired song in RBT task. According to the segment boundary information of these keywords, the tone scores of N-best hypotheses were computed by the MLP tone recognizer. Finally, the integration of tone score and acoustic score was then used to rescore the N-best hypotheses.

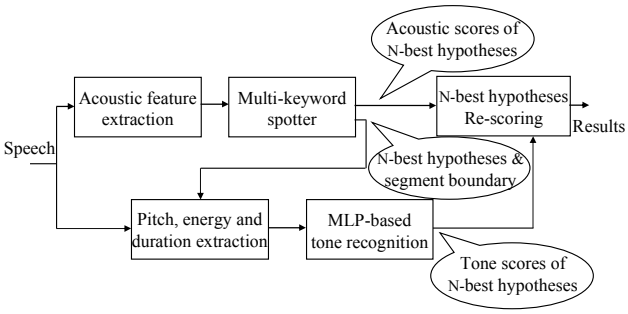

Figure 3: A functional block diagram of the proposed scheme.

The integration score of each hypothesis can be modeled by the weighted sum of acoustic score and tone score.

$$IS_{h_i} = AS_{h_i} \times w_a + TS_{h_i} \times w_t \quad (1)$$

where IS_{h_i}, AS_{h_i} and TS_{h_i} are the integration score, acoustic score and tone score of *i*-th hypothesis, respectively. The two weights w_a and w_t stand for the corresponding weights of acoustic score and tone score. The hypothesis-based tone score, TS_{h_i}, can then be expressed by the arithmetic mean of word-based tone scores, i.e.

$$TS_{h_i} = \frac{\sum_{j=0}^{NW_{h_i}-1} TS_{w_j}}{NW_{h_i}} \quad (2)$$

where NW_{h_i} is the number of words of *i*-th hypothesis and TS_{w_j} stands for the tone score of *j*-th word in the hypothesis. The word-based tone score, TS_{w_j}, can be further defined by the power mean of syllable-based tone scores, i.e.

$$TS_{w_j} = \left(\frac{\sum_{k=0}^{NS_{w_j}-1} \left(TS_{s_k}\right)^{\lambda}}{NS_{w_j}} \right)^{\frac{1}{\lambda}} \quad (3)$$

where NS_{w_j} is the number of syllables of *j*-th word in the hypothesis and TS_{s_k} stands for the tone score of *k*-th syllable in the word. The syllable-based tone score, TS_{s_k}, with value ranging between 0 and 1 is generated by the basic MLP tone recognizer. The parameters, w_a, w_t, and λ in above equations are determined based on some pilot experiments and set to be 1, 0.028 and 0.1, respectively. The usage of w_a and w_t, is to compensate the difference between the MLP and HMM outputs, while the reason of choosing lower value for λ is to emphasize the importance of syllables with lower tone score.

3.3: EXPERIMENTAL RESULTS OF INTEGRATION OF TONE RECOGNITION AND CONTINUE SPEECH RECOGNITION

The database used in this experiment consists of two sets. The first set is same as that used in tone modeling as described in Section 2.4. The second set is collected from the spontaneous inquiries of mobile RBT service. The first set is used for training only, while the second set is divided into two parts. The one containing 20,150 utterances (about 105,000 syllables) was used for training and the other containing 1933 utterances (about 11,000 syllables) was used for testing. All the utterances were collected over the GSM telephone network.

The top-1 and top-3 inclusion rate of RBT system have been shown in Table 2. It is clear that the integration of tone recognition and continuous speech recognition can efficiently improve the performance of RBT system.

Table 2: The results of continuous speech recognition with and without tone recognition (TR)

	Top 1 (%)	Top 3 (%)
Without TR	80.7	89.3
With TR	85.7	92.2

4: CONCLUSIONS

Tone recognition is critical to human perception of natural speech, especially for tonal language. How to use it as an additional knowledge source for automatic speech recognition is a challenging problem. In this paper, some MLP based tone recognition schemes have been proposed. Experimental results have confirmed that the effectiveness of both basic MLP and modular MLP schemes. Furthermore, under the framework of large vocabulary continuous speech recognition, we study the possible ways to incorporate tone recognition into speech recognition. The experimental results show that the integration of acoustic score and tone score is helpful to improve the overall performance of continuous speech recognition in the mobile RBT system.

REFERENCES

[1] Cheng, C. C., *A Synchronic Phonology of Mandarin Chinese*, Mouton, The Hague, 1973.

[2] Huang, H., Seide, F., "Pitch Tracking and Tone Features for Mandarin Speech Recognition", In *Proceedings of ICASSP*, pp. 1523-1526, 2000.

[3] Wong, Y. H., Chang, E., "The Effect of Pitch and Tone on Different Mandarin Speech Recognition Tasks", In *Proceedings of the 7th EUROSPEECH*, pp. 1517-1521, 2001.

[4] Qian, Y., Lee, T., and Soong, F., "Use of Tone Information in Continuous Cantonese Speech Recognition", In *Proceedings of International Conference on Speech Prosody*, pp. 587-590, 2004.

[5] Lin, C. H., Wu, C. H., Ting, P. Y., and Wang, H. M., "Frameworks for Recognition of Mandarin Syllables with Tone Using Sub-syllabic Units", Speech Communication, vol. 18, no. 2, pp. 175-190, 1996.

[6] Um, Ig-Tae, Won, Jong-Jin, and Kim, Moon-Hyun, "Text Independent Speaker Verification Using Modular Neural Network", In *Proceedings of IJCNN*, vol. 6, pp. 97-102, 2000.

[7] Lucas, S., Zhao, Z., Cawley, G., and Noakes, P., "On Decomposing MLPs", In *Proceedings of IEEE International Conference on Neural Networks*, vol. 3, pp. 1414-1418, 1993.