Improved Large Vocabulary Mandarin Speech Recognition Using Prosodic Features

Jui-Ting Huang & Lin-shan Lee

Graduate Institute of Communication Engineering
National Taiwan University, Taipei, Taiwan
fororgan@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

Abstract

This paper presents a new framework for improved large vocabulary Mandarin speech recognition using prosodic features. The prosodic information is formulated in a probabilistic model well compatible to the conventional maximum a posteriori (MAP) framework for large vocabulary speech recognition. A set of prosodic features considering the special characteristics of Mandarin Chinese is developed, and both syllable-level and prosodic-word-level prosodic models are trained with the decision tree algorithm. A two-pass recognition process is used, in which each word arc in the word graph output by the first pass is rescored in the second pass using the two prosodic models. The experiments show the reasonable improvements in recognition accuracy. This approach does NOT require a prosodic labeled training corpus, and works for the large-scale speaker-independent task.

1. Introduction

Substantial efforts have been made in analyzing the prosody in natural human speech[1][2]. The issue of incorporating the prosodic information into speech recognition processes has emerged in recent years. Such information as phone durations[3], phrase boundaries[4][5] and accentual types[4] were shown to offer relatively limited but actually significant improvements in English word recognition accuracy. Some recent work on Italian language[6] focused on more restricted tasks rather than the general purpose task, and more obvious improvements were found. For Mandarin Chinese, RNN-based prosodic modeling approaches were proposed to detect word-boundary cues to be used in the decoding process[7], and some improvements in character accuracy were also obtained. However, the experiment for this approach was performed only for a speaker-dependent recognition task with a single-speaker corpus, therefore the problem of wide variety of prosodic behavior for different speakers in more realistic tasks was not considered yet. In addition, very often the various approaches of incorporating prosodic information in speech recognition require a training corpus with prosodic labels marked by human experts, whose cost is high if not prohibitive. In this paper, we propose a new approach to incorporate prosodic information in large vocabulary Mandarin speech recognition, which not only works for the large-scale speaker-independent task, but requires only a training corpus which does not include any additional prosodic labeling.

Tseng et al[1] constructed a hierarchical multi-layer prosody model for fluent Mandarin speech, in which each hierarchical layer considers different level of perceived prosodic entity. In this model, the layered nodes from the bottom up are the syllables, prosodic words, prosodic phrases or utterances, breath group and prosodic phrase groups. In the approach proposed in this paper, only the syllable-level and prosodic-word-level features are analyzed and incorporated into the speech recognition process.

In the following, section 2 presents the proposed approach of using prosodic features in the recognition process. Section 3 describes the experiments and the results. The conclusion is made in section 4.

2. Proposed Approach

2.1. Recognition with prosodic modeling

The conventional formula for speech recognition based on the maximum a posterior (MAP) principle is

$$W^* = \arg\max_W P(W \mid X) = \arg\max_W P(W) P(X \mid W), \quad (1)$$

where the word sequence $W=\{w_1, w_2, \ldots, w_n\}$ is composed of n lexical words, and w_j is the j-th lexical word. For a given sequence of acoustic feature vectors X, equation (1) indicates that the recognized result is the word sequence W* that maximizes the posterior probability $P(W|X)$, which can be decomposed into two parts: $P(W)$ contributed by the language model and $P(X|W)$ by the acoustic models.

Now assume we are also given a sequence of prosodic feature vectors $F=\{f_1, f_2, \ldots, f_n\}$, where f_j is the feature vector for the lexical word w_j, equation (1) can then be modified to

$$W^* = \arg\max_W P(W \mid X, F) = \arg\max_W P(W) P(X, F \mid W). \quad (2)$$

Here both X and F are observable while in equation (1) only X is observable. Now equation (2) can be rewritten as

$$W^* \cong \arg\max_W P(W) P(X \mid W) P(F \mid W), \quad (3)$$

which is based on the assumption that the acoustic and prosodic feature sequences X and F are independent given the word sequence W. Here the first two terms are obtained from the same acoustic and language models as in equation (1), while the last term $P(F|W)$ is the probability obtained with the new prosodic model proposed here. If we assume that each of the prosodic feature vector f_j behaves independently given the word sequence, and in addition that only the current lexical word w_j have influence on its corresponding prosodic feature vector f_j., equation (3) becomes

$$W^* \cong \arg\max_W P(W) P(X \mid W) \prod_{j=1}^{N} P(f_j \mid w_j). \quad (4)$$

Here the assumptions are apparently not exactly correct, but can simplify the problem here in the initial studies. They can be modified to include more general conditions to be used in equation (4) in the future.

The above formulation with equations (2)-(4) can be used in either a one-pass decoding process or the rescoring stage in a multi-pass decoding process. The latter case is used here in this paper, in which we incorporate the prosodic information in the rescoring stage of a two-pass recognition process. The block diagram of the complete recognition process is depicted in Figure 1. For each input speech utterance, the first pass produces a word graph of a suitable size using a baseline recognition system with the conventional acoustic and language models producing $P(X|W)$ and $P(W)$. The second pass then rescores every word arc with lexical word hypothesis w_j in the graph by incorporating the prosodic model score $P(f_j|w_j)$ and $P(F|W)$ as specified in equations (3)(4), where the detailed evaluation of $P(f_j|w_j)$ will be given below. In Figure 1, every lexical word hypothesis w_j is composed of a few syllables represented by square blocks, and the prosodic model producing the probabilities $P(f_j|w_j)$ and $P(F|W)$ includes a set of decision trees, as will be clear below. The rescoring formula is then

$$S(W) = P(X|W) + \lambda_l P(W) + \lambda_p P(F|W), \tag{5}$$

where W is the word sequence hypothesis in the word graph being considered, λ_l, λ_p are the weighting coefficients for the language and prosodic model likelihoods with respect to the acoustic model likelihood, and $S(W)$ is the final score.

2.2. Prosodic modeling

The approach summarized above can be equally applied to all different languages, but the part below to formulating the probability $P(f_j|w_j)$ to be used in equation (4) is specifically for Mandarin Chinese. Chinese language is monosyllable-based. Every character has its own meaning and is pronounced as a monosyllable. A lexical word is then composed of one to several characters or syllables. However, it is well known that the lexical word cannot be considered as the basic unit in the prosodic structure in Mandarin speech. Instead, natural utterances automatically arrange the applicable combination of characters into "prosodic words", which is then the real basic prosodic unit in Mandarin speech production and perception. Such "prosodic words" are very often different from the lexical words. These "prosodic words" apparently carry plenty of prosodic information helpful to speech recognition. Furthermore, Chinese is a tone language. Each syllable is assigned a tone. There are a total of five different tones in Mandarin Chinese, including four lexical tones plus one neutral tone. Considering the above special characteristics of Mandarin speech, we develop the prosodic model from two levels, the syllable level and the prosodic word level, as in the following, in order to evaluate the probability $P(f_j|w_j)$ to be used in equation (4).

2.2.1. Syllable-level modeling

The probability $P(f_j|w_j)$ to be used in equation (4) can be evaluated from the syllable level as follows.

$$P(f_j|w_j) = \prod_{k=1}^{L_j} P(f_{jk}|T_{jk}, B_{jk}), \tag{6}$$

where L_j is the length of the word w_j, or the number of characters (or syllables) in w_j, f_{jk} is the vector constructed by the prosodic features extracted for the boundary right after the k-th syllable of the lexical word w_j, T_{jk} is the tone of the k-th syllable of the lexical word w_j, and B_{jk} is a variable indicating the end of the word, i.e., B_{jk}=1 if k= L_j and B_{jk}=0 otherwise. The definitions for all these symbols are clearly shown in an example word sequence in Figure 2. We have f_j={f_{jk}, k=1,2,…, L_j}. Now the probability $P(f_j|w_j)$ can be evaluated from the feature vectors for all the syllable boundaries of the lexical word hypothesis w_j as in equation (6).

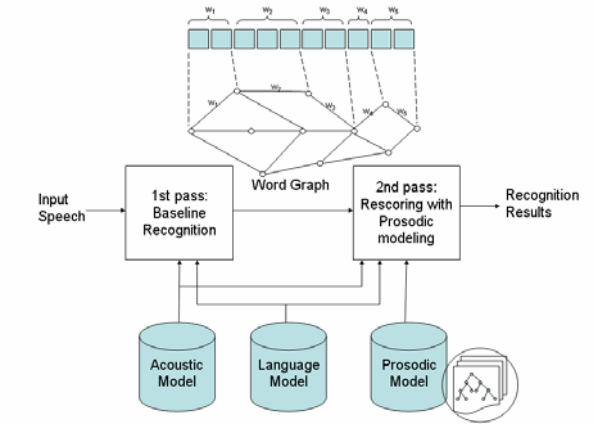

Figure 1: *The complete recognition process and rescoring, where each hypothesis lexical word arc is composed of a dew syllables represented by square blocks, and the prosodic model includes a set of decision*

By Bayesian theorem, we can have

$$P(f_{ik}|T_{ik}, B_{ik}) = \frac{P(T_{ik}|f_{ik}, B_{ik})P(f_{ik}, B_{ik})}{P(T_{ik}, B_{ik})}. \tag{7}$$

For a given lexical word hypothesis in the word graph, the component syllables are all given and thus the corresponding tone and boundary are already determined. It is therefore reasonable to set $P(T_{jk},B_{jk})$ to be unity without loss of generality. On the other hand, although f_{jk} is given for each syllable boundary, B_{jk} may be different for different lexical word hypothesis across the same syllable boundary. But here we set $P(f_{jk},B_{jk})$ to be a constant for lack of knowledge, because the approach to evaluate this probability is unknown at the moment. With the above, the probability $P(f_{jk}|T_{jk},B_{jk})$ in equation (6) will be estimated using the probability $P(T_{jk}|f_{jk},B_{jk})$ in equation (7), which can be obtained by decision trees as explained in section 2.4.

2.2.2. Prosodic-word-level modeling

There are in general three types of relations between the lexical words and the prosodic words in Chinese: (1) a prosodic word is a lexical word, (2) a prosodic word is a combination of several short lexical words, and (3) a prosodic word is a part of a long lexical word. Case (3) is very unusual, so we focus on the first two cases, in which we assume a prosodic word boundary never exists within a lexical word. In this case $P(f_j|w_j)$ in equation (4) can be evaluated on the prosodic word level by

$$P_2(f_j|w_j) = \begin{cases} \prod_{k=1}^{L_j-1} P(f_{jk}|B_{jk}), & \text{if } L_j \geq 2 \\ \text{a given constant}, & \text{if } L_j = 1 \end{cases}, \tag{8}$$

where B_{jk} is the same as defined previously. Here in the upper formula we assume all syllable boundaries which are not lexical word boundaries ($k \leq L_j-1$ for $L_j \geq 2$ as in equation (8)) are not prosodic word boundaries either. The mono-syllabic lexical word (L_j=1 in the lower formula in equation (8)), on the other hand, is usually connected to the preceding or following lexical word to form a prosodic word. The situation for this case is different to model at the moment, so we only set a given constant here. By Bayesian theorem,

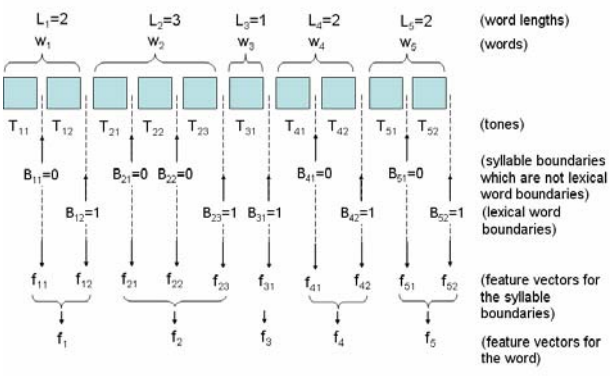

Figure 2 : *The definition of the symbols wj, Lj, Tjk, Bjk, fjk, used in equation (4) and (6) for an example word sequence hypothesis W={w1, w2, w3, w4, w5}. Here every square block represents a syllable. For example, w1 has two syllables, and so on.*

$$P\left(f_{jk} \mid B_{jk}\right)=\frac{P\left(B_{jk} \mid f_{jk}\right)P\left(f_{jk}\right)}{P\left(B_{jk}\right)}, \tag{9}$$

same as equation (7), $P(B_{jk})$ is set to unity for a given lexical word hypothesis w_j in the word graph, and $P(f_{jk})$ is to be a constant value for lack of knowledge although it should be different for different lexical word hypothesis w_j across the same syllable boundary. So the probability $P(f_{jk}|B_{jk})$ in equation (8) will be estimated using the probability $P(B_{jk}|f_{jk})$ in equation (9), which can be obtained by decision trees as explained in section 2.4..

2.2.3. Integrated prosodic modeling

In order to use both levels of information as obtained above, weighted sum of the two probabilities $P_1(f_j|w_j)$ and $P_2(f_j|w_j)$ obtained from equation (6) and (8) respectively is used:

$$P\left(f_{\mathrm{j}} \mid w_{\mathrm{j}}\right)=\alpha P_1\left(f_{\mathrm{j}} \mid w_{\mathrm{j}}\right)+(1-\alpha) P_2\left(f_{\mathrm{j}} \mid w_{\mathrm{j}}\right), \tag{10}$$

where α ranges from 0 to 1.

2.3. Feature Parameters

As mentioned above, a set of features are extracted for each syllable boundary. These features can be divided into two types, the prosodic features and the categorical features. The categorical features were not mentioned above, but they are helpful and can be easily inserted into the feature vectors f_{jk} and f_j, as will be explained below.

2.3.1. Prosodic features

Prosodic features are derived from pitch, duration and energy. Pitch has been proved very useful in the recognition of the tone for Mandarin Chinese, and here we use it to derive as many different pitch-related features as possible. For each syllable boundary detected, various pitch-related features were calculated using the pitch contour within the two syllables right before and right after the boundary. The example features used include the average value of the pitch within the syllable, the average of the absolute value of the pitch slope within the syllable, the range of the pitch within the syllable, the pitch reset across the boundary, and so on. In order to represent the shape of the pitch contour within a syllable, we also used the first four coefficients of the Legendre discrete polynomial expansion of the contour[8], for which the zero-th order coefficient represents the level of the contour, and the other three coefficients represent the key characteristics of the contour shape. A total of 16 pitch-related attributes were used here for each syllable boundary.

Duration features such as pause and phone durations have been used to describe the phenomena of the prosodic continuity and the pre-boundary lengthening[9]. The durations of the two syllables before the boundary being considered and the ratio of them are example features used here. Energy also has similar effects on the prosodic structure[1]. The average and peak energy of the syllable before and after the boundary as well as the ratio of them are example features used here. Finally, a total of 8 duration- and energy-related features were used.

2.3.2. Categorical features

A Mandarin syllable is conventionally decomposed into two parts: the Initial and the Final. Here Initial is the initial consonant part of the syllable, while the Final includes the vowel part plus the optional medial and ending. Since the types of both the Initial and the Final for a syllable apparently influences the pitch, duration, and energy features, we created 59 binary attributes corresponding to each of the 37 Final and 22 Initial types respectively. Only the two attributes for the Initial/Final of the syllable right before the boundary being considered were set to one, and all the others to zero. These attributes provide the information regarding which Initial/Final the prosodic features obtained above belong to.

Prosodic features apparently behave depending on the tones of the syllables being considered. Therefore, we created another total of 10 binary attributes, each corresponding to the five tones of the two syllables on both sides of the syllable being considered. Only two attributes for the tones of the two syllables mentioned above were set to one and the others to zero.

2.4. Prosodic model training with decision trees

The decision tree algorithm was used for the prosodic model to estimate the probabilities $P(T_{jk}|f_{jk},B_{jk})$ in the syllable-level model (equation (7)) and $P(B_{jk}|f_{jk})$ in prosodic-word-level model (equation (9)). In the case of syllable-level model, we divided the training prosodic feature vectors into two cases: B_{jk}=1 or 0, or lexical word end or not. We then trained two decision tree models for each case respectively, one to estimate the probability $P(T_{jk}|f_{jk},B_{jk}=0)$ and the other $P(T_{jk}|f_{jk},B_{jk}=1)$. For the prosodic-word-level model, only one decision tree model is needed to estimate the probability $P(B_{jk}|f_{jk})$.

The RandomForest package in the software R[10] was selected in the implementation of the decision trees. In the training phase, many decision trees were trained for each case, in which m out of M variables were randomly selected at each node, where M is the total number of variables in the feature vectors, and $m<<M$. The best split based on these m variables is then used to split the node. In the testing phase, for each input feature vector, each tree gives a classification, and the final classification is based on the "votes" of the trees. The probability for each class is then obtained by the percentage of the votes for that class.

To prepare the training data for the prosodic-word-level model, we first deleted the feature vectors for the boundaries on both sides of the mono-syllabic lexical words, because as mentioned previously such words are always connected with the neighboring lexical words to form a prosodic word, and such situation is different to model at the moment. We then

Table 1: *The top six important prosodic features for the two prosodic models*

rank	Syllable-level model	Prosodic-word-level model
1	the average pitch slope	the reset of the average energy across the boundary
2	the pitch slope at the beginning 3 frames of the syllable after the boundary	the pitch at the beginning frame of the syllable after the boundary
3	the third Legendre coefficient	the average energy over the syllable after the boundary
4	the pitch at the beginning frame of the syllable after the boundary	the third Legendre coefficient
5	the second Legendre coefficient	the average energy over the syllable before the boundary
6	the duration of the syllable before the boundary	the reset of the pitch across the boundary

Table 2: *Rescoring results: character accuracy(%)*

baseline	Rescoring with			
	λ_p=7.0	λ_p=8.0	λ_p=9.0	λ_p=10.0
80.78	81.72	81.75	**81.84**	81.79

train the model of decision trees using the rest of boundaries based on B_{jk}=0 or 1.

3. Experimental Results

3.1. Corpus and experimental setup

The corpus used in this research was taken from the Chinese Broadcast News Corpus (CBN), which was recorded from a few radio stations in Taipei in 2001. The corpus used here include a total of 9806 utterances (10 hours) produced by nine female and five male speakers, all with the correct text transcription. 8731 utterances out of them were used for training, while the rest 1075 utterances for testing. All the speakers distribute equally on both the training and testing sets. The recognition experiments were performed with a lexicon of 100K entries, a trigram language model, and an intra-syllable right context dependent Initial/Final acoustic model set.

3.2. Importance of the features

We first examine the importance of each individual prosodic feature used here. In the RandomForest-type decision tree, the split of a node is made when the gini impurity metric (similar to the entropy) for the two child nodes is less than that of the parent node. Therefore, adding up the reduction in the gini impurity for the features over all trees gives a fast estimate of the importance of the features. Table 1 lists the top six important features for the syllable and prosodic-word level models respectively. We found the important features are significantly different in the two models. For the syllable-level model, all important features are pitch-related, apparently because the target label to be classified here is the tone, i.e., $P(T_{jk}|f_{jk},B_{jk})$ in equation (7) is used in the classification.

For the prosodic-word-level model, the energy-related features, especially the energy reset (rank 1 in Table 1), has the largest discriminative power for the prosodic word boundaries. This is consistent with the phenomenon that energy has the largest value in the beginning of any kind of prosodic unit and then attenuates gradually [1]. The pitch at the beginning frame (rank 2) and the pitch reset (rank 6) both have significant discriminative power, which implies that there is usually some prosodic continuity within the prosodic word or discontinuity between the prosodic words. Also, the third Legendre coefficient (rank 4) indicates that the pitch shape characteristics are related to the prosodic word identification as well.

3.3. Final recognition results

Table 2 lists the final recognition results after rescoring. The best result of character accuracy achievable here is 81.84% for λ_p=9.0 as compared to the baseline of 80.78%, which represents an error rate reduction of 5.5%. Notice that the upper bound character accuracy or the inclusion percentage for the word graph is 88.66%, which indicates that there is still plenty of space for further improvements. Also note that here a single model was used for all different speakers. Much better improvements may be possible if a different model for each speaker can be trained.

4. Conclusions

In this paper we propose a new approach for improving Mandarin speech recognition by incorporating the prosodic information. A new set of prosodic features were developed considering the special characteristics of the Chinese language, and two levels of prosodic model were trained using decision trees to generate the prosodic likelihood score to be used in the rescoring stage of the recognition process. The experiments performed on broadcast news corpus with many speakers verified the benefit of incorporating the prosody information in the recognizer, and the results are analyzed and discussed.

5. References

[1] Tseng, Chiu-yu et al. Fluent speech prosody: framework and modeling. *Speech Communication*, Vol.46, 284-309.

[2] H. Fujisaki. Information, prosody, and modeling—with emphasis on tonal features of speech. In *SP-2004*, 1-10.

[3] D. Vergyri et al. Prosodic knowledge source for automatic speech recognition. In *Proc. ICASSP*, vol. 1, pp.208-211, Hong Kong, 2003.

[4] K. Chen et al. Prosody dependent speech recognition with explicit duration modeling at intonational phrase boundaries. In *Proc. EUROSPEECH*, pp.393-396, GENEVA, 2003

[5] A. Stolcke et al. Modeling the prosody of hidden events for improved word recognition. In *Proc. EUROSPEECH*, vol. 1, pp.307-310, Budapest, 1999.

[6] R. Gretter et al. Using prosodic information for disambiguation purposes. In *Proc. EUROSPEECH*, pp.1821-1824, Lisboa, 2005.

[7] W. J. Wang et al. RNN-based prosodic modeling for mandarin speech and its application to speech-to-text conversion. *Speech Communication*, Vol.36, 247-265.

[8] S. H. Chen et al. Vector Quantization of Pitch Information in Mandarin Speech. IEEE trans. On Communications, 38(9), 1317-1320.

[9] Shriberg, E. et al. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 2000, pp.127-154.

[10] http://stat.www.berkeley.edu/users/breiman/RandomForests