Prosodic Modeling in Large Vocabulary Mandarin Speech Recognition

Jui-Ting Huang and Lin-shan Lee

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

ABSTRACT

The issue of incorporating prosodic information into speech recognition processes has emerged in recent years. In this work we present a complete framework for Mandarin speech recognition with prosodic modeling considering two-level hierarchical prosodic information for Mandarin Chinese. We developed a GMM-based, a decision-tree-based, and a hybrid approach. The best improvements in character recognition accuracy were obtained by the decision-tree-based prosodic models. This approach does NOT require a training corpus labeled with prosodic features, and works reasonably for a large-scale multi-speaker task.

Index Terms: prosody, speech recognition, Mandarin tone, lexical word boundaries

1. INTRODUCTION

It is well known that the prosody of speech carries plenty of linguistic information, but it is difficult to use such information in speech recognition. Some recent studies tried to incorporate into the prosody such as phone durations [1], phrase boundaries [2,3] and accentual types [2] in speech recognition, but relatively limited improvements were achieved in recognition accuracy. Some researchers [4] analyzed carefully the different ways to construct accent type/tone type models for Japanese/Chinese, to control the pruning size in decoding, and to construct language models including prosodic events. However, 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 addition, there is also the problem of the wide variety of prosodic behavior for different speakers in more realistic tasks.

Recently we have shown that [5] the word-boundary indicator not only has to do with prosodic features, but is helpful to recognition as well. In this paper, we take into account the recently proposed hierarchical prosody framework for Mandarin Chinese [6] to integrate both syllable- and word-level information with hierarchical relationships to be used in large vocabulary speech recognition. The approach proposed in this paper is also towards the goal of a large-scale speaker-independent task which requires only training corpora without prosodic labeling. Reasonable improvements in character recognition accuracy were obtained.

In the following, the recognition framework, including prosodic features, are presented in Section 2, while Section 3 discusses the three approaches for prosodic modeling proposed here. The experimental results are then presented in section 4. The conclusion is made in section 5.

2. RECOGNITION FRAMEWORK

2.1 Overall picture

In contrast to the conventional approach for speech recognition, and in addition to the observable sequence of acoustic feature vectors X, here it is assumed that an additional observable sequence of prosodic feature vectors F is also available. Therefore the recognition formula based on the maximum *a posterior* (MAP) principle is

$$W^* = \arg\max_W P(W \mid X, F) \tag{1}$$

$$= \arg\max_W P(W) P(X, F \mid W), \tag{2}$$

where the word sequence $W=\{w_1, w_2, ..., w_N\}$ is composed of N lexical words, and w_j is the j-th lexical word; $F=\{f_1, f_2, ..., f_N\}$ is composed of N prosodic feature vectors each for a lexical word, and f_j is the prosodic feature vector for the lexical word w_j. The recognition result is the word sequence W^* that maximizes the posterior probability $P(W|X,F)$, which can be decomposed into two parts: $P(W)$ and $P(X,F|W)$. If we assume that the acoustic and prosodic feature sequences, X and F, are independent given the word sequence W (which may not be true,) equation (2) can be written as

$$W^* \cong \arg\max_W P(W) P(X \mid W) P(F \mid W). \tag{3}$$

Here $P(W)$ is contributed by the language model and $P(X|W)$ by the acoustic model, while the last term $P(F|W)$ is the probability obtained with the prosodic modeling proposed here and presented in this paper.

The recognition proposed here is based on a two-pass process as shown in Figure 1. For each input speech utterance, the first pass produces a word lattice using a baseline recognition system with the conventional acoustic and language models. The second pass then rescores every word arc in the lattice by incorporating the prosodic model score. The rescoring formula is directly obtained from equation (3),

$$S(W) = \log P(X \mid W) + \lambda_l \log P(W) + \lambda_p \log P(F \mid W), \tag{4}$$

where W is the word sequence hypothesis in the word graph

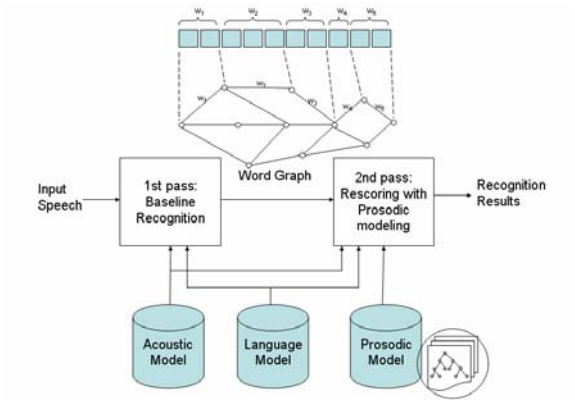

Figure 1. The complete recognition process and rescoring, where each hypothesis lexical word arc is composed of a dew syllables represented by square blocks, and the prosodic model includes a set of decision

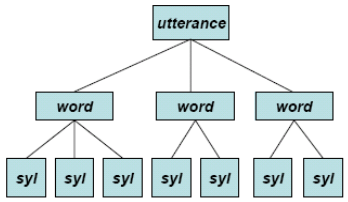

Figure 2. The basic structure for Mandarin speech recognition. An utterance can be divided into several lexical words, and each lexical word consists of one to several syllables.

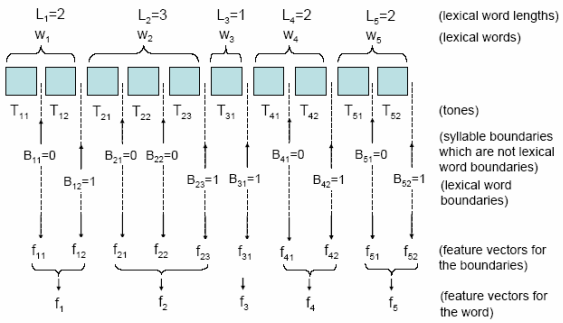

Figure 3. The definition of the symbols w_j, L_j, T_{jk}, B_{jk}, f_{jk} used in equation (4) and (6) for an example word sequence hypothesis W={w1, w2, w3, w4, w5}. Here every square block represents a syllable. For example, w1 has two syllables, and so on.

being considered; λ_l and λ_p are respectively the weighting coefficients for the language and prosodic model scores with respect to the acoustic model likelihood; and $S(W)$ is the final score. After rescoring every possible patch, the path with the highest score is the recognition result.

2.2 Mandarin speech recognition

The prosodic model proposed here is based on the special characteristics of Mandarin Chinese. Chinese language is monosyllable-based. A lexical word is composed of one to several characters, and each character is pronounced as a monosyllable. Chinese is also 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. We take into account the above characteristics as well as the hierarchical framework of speech prosody [6] to come up with a basic structure for Mandarin speech recognition as illustrated in Figure 2. An utterance can be divided into several lexical words, and each lexical word consists of one to several syllables. At the syllable level, the tone type is the major influential factor for prosodic behavior. We define a random variable $T = t,\ t \in \{1,2,3,4,5\}$ for every syllable, where the number indicates the tone type, and the value of 5 refers to the neutral tone. The lexical word level is above the syllable level, in which the locations of the syllables and the lexical word boundary apparently influence the prosodic behavior. We hence define the lexical word boundary indicator as a random variable $B = b,\ b \in \{0,1\}$ for each syllable boundary. B=1 indicates that the syllable boundary is a lexical word boundary, while B=0 indicates that it is not.

Now consider the prosodic likelihood $P(F|W)$ for prosodic modeling as mentioned above. Under the assumption that each of the prosodic feature vectors f_j for each word w_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 (these assumptions are certainly not true, but lead to simplified modeling and approximated solutions), we can write

$$P(F|W) = \prod_{j=1}^{N} P(f_j | w_j). \tag{5}$$

Because a lexical word consists of one to several syllables, we may take this word-level likelihood as the product of all syllable-level likelihoods and represent the given syllable by its tone type and the lexical word boundary indicator (again under some not really true assumptions):

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

where L_j is the length of the lexical 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 random variable T previously defined for the tone of the k-th syllable of the lexical word w_j; and B_{jk} is a random variable B previously defined for the lexical word boundary indicator 0 or 1 for the boundary after the k-th syllable of the lexical word w_j. The definitions for all these symbols are clearly shown in an example word sequence W in Figure 3.

2.3 Feature parameters

As mentioned above, a set of features were extracted for each syllable boundary for the word lattice obtained in the first pass to form a feature vector f_{jk} for the boundary after the k-th syllable of the lexical word w_j. 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 for decision tree training introduced in section 3. Details about the categorical features are discussed in [5].

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 [7], 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 [8]. 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 [6]. 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.

3. PROSODIC MODELS

This section investigates the modeling approaches used to estimate the probability $P(f_{jk}|T_{jk},B_{jk})$ in equation (6). A total of three approaches were developed, which are presented below.

3.1 GMM-based approach

The GMM-based approach is very straightforward. We directly model the probability $P(f_{jk}|T_{jk},B_{jk})$ using a Gaussian mixture model (GMM). Each pair of $(T,B)=(t,b)$ is considered as a class, so there are a total of nine classes: $(T,B)= (t,b)$, $(t,b) \in \{(1,0), (2,0), (3,0), (4,0), (1,1), (2,1), \ldots, (5,1)\}$ for the five tone types and two lexical word boundary conditions. Here we excluded the pair of (t,b)=(5,0) because the neutral tone is never observed in the middle of a word. We then train a GMM for the prosodic features, as shown in Figure 4(a) to describe their distribution for each of the above nine classes. This modeling is analogous to the conventional acoustic modeling, in which we train a Hidden Markov Model (HMM) to describe the distribution of the acoustic features like MFCC for each phone class. The training is based on the maximum likelihood estimate (MLE) criterion, also similar to acoustic modeling. This model optimization criterion can avoid the computation of all rival classes, but at the cost of a relatively weak discriminative power [9].

3.2 Decision-tree-based approach

According to Baye's rule, $P(f_{jk}|T_{jk},B_{jk})$ can be written as:

$$P\left(f_{jk}\left|T_{jk},B_{jk}\right.\right)=\frac{P\left(T_{jk},B_{jk}\left|f_{jk}\right.\right)P\left(f_{jk}\right)}{P\left(T_{jk},B_{jk}\right)}. \quad (7)$$

The first probability $P(T_{jk},B_{jk}|f_{jk})$ in the numerator is the local *a posteriori* probability, which can be decomposed into two parts:

$$P\left(T_{jk},B_{jk}\left|f_{jk}\right.\right)=P\left(B_{jk}\left|f_{jk}\right.\right)P\left(T_{jk}\left|f_{jk},B_{jk}\right.\right). \quad (8)$$

We therefore trained two decision trees to estimate respectively the two probabilities $P(B_{jk}|f_{jk})$ and $P(T_{jk}|f_{jk},B_{jk})$. The first probability $P(B_{jk}|f_{jk})$ carries lexical-word-level prosodic information because it presents the probability of a syllable being a word-ending syllable given the prosodic features. The probability $P(T_{jk}|f_{jk},B_{jk})$, on the other hand, carries the syllable-level modeling because it represents the probability of a syllable being of a certain tone type given the prosodic features and the lexical word boundary information. We can see that the syllable-level modeling actually depends on the lexical-word-level information, which means the prosodic modeling here is in fact hierarchical. The probability $P(T_{jk},B_{jk})$ in the denominator is the prior probability, which can be estimated based on the training data simply by counting the number of times that each class of (t,b) pair appears in the training set. The probability $P(f_{jk})$ is slightly difficult to estimate, although the features can be easily extracted from the syllable boundary. This is because the same boundary may have many different (j,k) indices for the k-th syllable boundary in the lexical word w_j for different paths in the word graph across the boundary. We thus simplified the problem by assuming $P(f_{jk})$ to be a constant.

3.3 Hybrid approach

Although the decision-tree-based approach can contribute more to the discriminative prosodic score, it is based on a simplified assumption that $P(f_{jk})$ is constant among different paths. The hybrid approach proposed here is to try to keep the advantages of the decision tree estimates but eliminating its simplified assumption by merging the above two approaches. For the prosodic feature f_{jk}, the output of the decision tree can be seen as a vector composed of nine posterior probabilities $P(T_{jk},B_{jk}| f_{jk})$ corresponding to the nine classes of (t,b) pairs. Here we denote it as a vector $f_{jk}^{(p)}$. Now instead of using prosodic features f_{jk}, we used this vector of nine posterior probabilities, $f_{jk}^{(p)}$, as the "feature" inputs for GMM. In other words, the output probability of GMM becomes $P(f_{jk}^{(p)}|T_{jk},B_{jk})$ instead of $P(f_{jk}|T_{jk},B_{jk})$. We also try another possible feature input for GMM by concatenating $f_{jk}^{(p)}$ with f_{jk}, that is, the output probability becomes $P(f_{jk}, f_{jk}^{(p)}|T_{jk},B_{jk})$.

Table 1. Character accuracies and classification rates for different approaches. Hybrid 1 indicates that only the class probabilities are used; Hybrid 2 indicates that the class probabilities are concatenated with the original prosodic feature vector.

	(A) Classification rate (%)	(B)Character accuracy after rescoring (%)	(C)=(B)-(A) Character accuracy improvements(%)
Baseline	-----	80.78	-----
GMM-based	34.55	81.13	0.35
Decision-tree-based	62.77	82.23	1.45
Hybrid 1	51.62	81.88	1.10
Hybrid 2	43.16	81.43	0.65

4. EXPERIMENTS

4.1 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 transcriptions. 8731 utterances out of them were used for training, while the rest (1075 utterances) were used for testing. The utterances of each of the speakers were distributed equally in 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.

4.2 Experimental results

Table 1 shows the character recognition accuracy for each approach after rescoring. Also listed in the table is the classification rate for each approach. In other words, given a prosodic feature f_{jk} for the k-th syllable boundary in the lexical word w_j, the classification into one of the nine (t,b) classes is based on the output probabilities—$P(f_{jk}|T_{jk},B_{jk})$ in the GMM-based approach, $P(T_{jk},B_{jk}|f_{jk})$ in the decision-tree-based approach, and $P(f_{jk}^{(p)}|T_{jk},B_{jk})$ or $P(f_{jk}, f_{jk}^{(p)}|T_{jk},B_{jk})$ in the hybrid 1 or hybrid 2 approach. GMM has the worse classification rate because we used all the prosodic features in GMM modeling without considering the importance or discriminative power of each feature; therefore some noisy features may inevitably cause some confusion. The decision tree has the best classification rate because the posterior probabilities $P(T_{jk},B_{jk}|f_{jk})$ used for classification are known to be the optimal . In the hybrid approach, using the class probabilities alone (hybrid 1) gives better results than concatenating it with the original prosodic features (hybrid 2), which implies that although the class probabilities have the discriminative power, the original prosodic features are somewhat confusing and thus degrading the discrimination capabilities as well.

The character accuracies, especially the improvements in the character accuracies in the last column of Table 1, are highly correlated with the classification rates. The baseline accuracy for the first pass recognition in the framework of Figure 1 is 80.78%. After incorporation with prosodic models, we found that all the three approaches offered improved accuracy. The improvements obtainable from the GMM alone seem to be limited, probably because this model is easily disturbed by noisy features. The decision-tree-based approach, on the other hand, is able to offer the best improvements, probably because it is able to identify the more discriminative features and overlook that are not helpful. Although the hybrid approach performs better than GMM-based approach, it cannot surpass the decision-tree-based one. Hybrid 1 offered some improvements based on the class probabilities only, so it is reasonable to expect more improvements if more features are included as in Hybrid 2. However, just as with the degraded classification rate for Hybrid 2, the recognition accuracy also became worse, probably due to the deficiency of the GMM that cannot handle the noisy features.

5. CONCLUSIONS

We propose a prosody-incorporated probabilistic framework for Mandarin speech recognition, in which the prosodic model is designed considering the syllable and word-level information. GMM, decision tree, and hybrid approaches are used respectively to construct the prosodic model, and all three methods can produce a reliable prosodic likelihood to be incorporated in the rescoring formula to improve the recognition accuracy. The decision-tree-based approach improved the character recognition accuracy the most.

6. REFERENCE

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

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

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

[4] K. Hirose & N. Minematsu. Use of Prosodic Features in Speech Recognition. *Proc. ICSLP*, 2004

[5] J-T Huang & L-S Lee. Improved Large Vocabulary Mandarin Speech using prosodic features. *Speech Prosody*, 2006.

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

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

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

[9] H. Bourlard & N. Morgan. Continuous Speech Recognition by Connectionist Statistical Methods. *IEEE trans. On Neural Networks,* 4(6), 893-909