Improved Large Vocabulary Mandarin Speech Recognition by Selectively Using Tone Information with a Two-stage Prosodic Model

Li-Wei Cheng and Lin-Shan Lee

Graduate Institute of Communication Engineering, National Taiwan University, Taiwan

R95942129@ntu.edu.tw, lslee@gate.sinica.edu.tw

Abstract

The incorporation of prosodic information in large vocabulary continuous speech recognition has attracted much attention in recent years, especially for a tonal language such as Mandarin Chinese. The tones of some syllables are very difficult to recognize correctly due to the very complicated prosodic behavior. Tone recognition errors inevitably degrade the recognition accuracy seriously. We propose a new approach by introducing an extra tone category of “unknown.” When the tone is difficult to recognize, the tone information will not be used. A two-stage prosodic model is developed for such a propose, and a 17.8% reduction in character error rate was achieved. Notably, this approach does not require speaker normalization for prosodic features.

Index Terms: speech recognition, Mandarin tone, prosody

1. Introduction

It is generally believed that the prosody of a native speaker carries much valuable linguistic information [1]. Recent work has attempted to take advantage of this information to improve recognition accuracy for large vocabulary continuous speech. Researchers have attempted to integrate prosodic information into speech recognition systems for tonal languages such as Cantonese [2,3] and Mandarin Chinese [4–7]. However, prosody of continuous speech in Mandarin is highly variable, even for the same speaker [8]. Improvements to recognition accuracy using current prosodic models are still not satisfactory. However, an upper bound of 32% error rate reduction (relative) was reported in [6] by rescoring with perfect tone information: this gives us hope that better ways may be found.

The most important factor in the prosody of tonal languages is the tone type. Many recent studies have focused on tone modeling. Some model pitch contour within the acoustic model [9], while others apply feature-based models for tone classification after first-pass acoustic alignment [10–12]. Resultant tone classification accuracies are typically impressive, but character accuracy results after incorporation into LVCSR are either omitted or show considerable room for improvement. For short song name inquiries [13] and small vocabulary tasks [14], tone recognition has proved helpful in improving character accuracy.

From a linguistic standpoint, the tones for 20 to 40 percent of the syllables in continuous Mandarin speech are difficult to recognize. That is, even for humans, many characters have tone patterns that are difficult to recognize without knowing the context[1]. As such, full accuracy in tone recognition may be an unreachable goal. Also, as general character recognition accuracy has approached or passed 80%, yielding word graphs with high inclusion rates, tone recognition is often used in two-pass rescoring [5,7]. Hence, it is not necessary to recognize tones for all characters. Moreover, in LVCSR, many hypothesized characters are misaligned, resulting in meaningless tone patterns. We thus propose a prosodic model that classifies characters as either a specific tone number, or as tone-unknown.

2. Proposed framework

2.1. Prosody-enhanced recognition

The conventional approach of speech recognition is to optimize the word sequence by maximizing the probability that the evidence yields the sequence given the acoustic feature vectors \(X\), such as MFCC. In our approach, the prosodic feature vectors \(F\) are integrated into the recognition formula. We apply the maximum a posteriori (MAP) principle as follows:

\[
W^* = \arg \max_{W} P(W|X, F) \\
\approx \arg \max_{W} P(W)P(X, F|W).
\]

where the lexicon word sequence \(W = \{W_1, W_2, \ldots, W_N\}\) is a time-order sequence of \(N\) lexicon words, each with one to several mono-syllable characters; \(F = \{F_1, F_2, \ldots, F_K\}\) is a set of prosodic feature vectors, while \(P_k\) is the set of syllable-related feature vectors for syllables in \(W_k\).

We assume that the prosodic feature vectors \(F\) are independent of the acoustic feature vectors \(X\) given the word sequence \(W\). Eq. 2 can be reformulated as

\[
W^* \approx \arg \max_{W} P(W)P(X|W)P(F|W).
\]

Hence, the recognition criteria is divided into three parts: the probability \(P(W)\) is obtained from the language model, \(P(X|W)\) is calculated from the acoustic model, and the last probability \(P(F|W)\) is contributed by the proposed prosodic model.

Then we assume the acoustic feature vectors are depended on current word, the prosodic feature vectors are depended on the preceding word and the current word, and the current word is depended on preceding word history. Eq. 3 can be reformulated as

\[
W^* \approx \arg \max_{W} \prod_{k} P(X_k|W_k) \\
\cdot \prod_{k} P(W_k|W_{k-1}, W_{k-2}, \ldots) \prod_{k} P(F_k|W_{k}, W_{k-1}),
\]

where \(W_k\) is the character sequence in arc \(k\), \(X_k\) is the acoustic feature vectors for \(W_k\), and \(F_k\) is the prosodic feature vectors for \(W_k\).

In the proposed approach, the extraction of prosodic feature vectors is dependent on syllable alignment. Therefore, the recognition process is based on a two-pass recognition evaluation. The first pass implements speech recognition with

---

1 Note that in Mandarin Chinese, which is a monosyllabic language, all characters are pronounced as mono-syllables.
the acoustic and language model, but instead of outputting the best word sequence hypothesis, it generates a word graph and passes it on to the second pass, in which the prosodic feature vectors are extracted using the word graph’s syllable alignments. According to Eq. 4, the second pass rescores the arcs in the word graph using the formula

$$S(W_k) = \log P(X_k|W_k) + \lambda_{LM} \log P(W_k|k-1, W_{k-2}, \ldots)$$

$$+ \lambda_{PM} \log P(F_k|W_k, W_{k-1})$$

(5)

where $S(\cdot)$ is the word arc scoring function. $\lambda_{LM}$ and $\lambda_{PM}$ are the weighting coefficients for the language and prosodic model scores with respect to the acoustic model score. In the end, the best path (word sequence hypothesis) is recalculated using the Viterbi algorithm. Fig. 1 illustrates the process of two-pass speech recognition.

Fig. 1. Two-pass speech recognition flow diagram.

2.2. Mandarin speech recognition with prosody

Our prosodic rescoring is based on characteristics of Mandarin Chinese, which is a monosyllabic and tonal language, where each character has a tone number. Since there are five different tones in Mandarin, specifying a tone for an unknown character can significantly reduce the size of the search space. The five tones include four lexical tones, {1, 2, 3, 4}, plus one neutral tone, {5}. These tone patterns are the most influential factor in the prosody synthesis of Mandarin utterances.

Consider the prosodic likelihood $P(F_k|W_k, W_{k-1})$ in Eq. 5. Let $T_k$ be the tone sequence for character sequence $W_k$ in arc $k$ of the word graph. By assuming tone sequences $T_k$ as the dominant factor in the generation of the prosodic feature vector set $F_k$, we write

$$P(F_k|W_k, W_{k-1}) = P(F_k|T_k, T_{k-1})$$

(6)

Generally, the current character’s tone and the preceding prosodic feature vector have the largest influence on the prosody of the current syllable. Assuming that the prosodic feature vector is decided solely by these two, we reformulate the probability $P(F_k|T_k, T_{k-1})$ as

$$P(F_k|T_k, T_{k-1}) = \prod_{j=1}^{N_k} P(f_j|t_j, t_{j-1}, f_{j-1}, f_{j-2}, \ldots)$$

(7)

$$\cong \prod_{j=1}^{N_k} P(f_j|t_j, f_{j-1})$$

(8)

where $N_k$ is the number of characters (syllables) in word graph arc $k$, $f_j$ is the prosodic feature vector of syllable $j$ (character $j$), and $t_j$ is the tone type of character $j$. Finally, taking the log likelihood of Eq. 8, we define the prosodic model score for word arc $k$ as

$$S_{PM}(W_k) = \sum_{j=1}^{N_k} \log P(f_j|t_j, f_{j-1}).$$

(9)

That is, the prosodic model score $S_{PM}(W_k)$ for Eq. 5 is decomposed into the sum of the character score $S_{PM}(w_j)$ of character $w_j$ in arc $k$.

2.3. Proposed prosodic model

We introduce the two-stage prosodic model used here, represented by the “prosodic model” square in Fig. 2. In the first stage, we use the extracted prosodic features to build a basic five tone model using a random forest, where a group of decision trees is grown into a forest, and the tone decision can be made by a majority vote. We, however, treat the random forest’s output as a distribution of votes which we output for next stage use. Also, we use another tone-unknown discriminator to assist in the second stage decision-making process. In the development of this tone-unknown discriminator, each syllable in the training set is labeled as either correct or incorrect using syllable tone information and the majority vote from the basic five tone model. Then, using these labels as training targets, we build another random forest classifier to identify tone-unknown syllables, i.e., those whose tone type is difficult to decide.

The second stage constructs a tone probability decision strategy. With the help of the tone-unknown syllable discriminator plus the distribution of tone votes from the basic five tone model, we determine whether the recognized syllable tone is reliable. Some syllables have been classified as tone-unknown by the discriminator in the first stage, but other syllables may still lack a majority voted tone from the basic five tone model. Hence, if the output distribution of the basic five tone model lacks a majority vote, we again viewed it as a tone-unknown syllable. All other syllables left are then taken as tone-reliable. Finally, for the output of the prosodic model, if the syllable is tone-reliable, we then use the voting result of the basic five tone model as the estimate of the probability distribution of the five tones. Otherwise, a low threshold probability is assigned to the output five probabilities of the five tones.

2.4. Prosodic model structure formulation

Here, we present the ways to estimate the probability $P(t_j|f_j, f_{j-1})$ required in Eq. 9. According to Bayes’ rule, the conditional probability in Eq. 9 can be written as

$$P(t_j|f_j, f_{j-1}) = \frac{P(t_j|f_{j-1})P(f_j|t_j, f_{j-1})}{P(f_j|t_j, f_{j-1})}.$$  (10)

The probability $P(t_j|f_{j-1})$ is estimated by the basic five tone model mentioned above. Although it is difficult to estimate the ratio $P(f_j|t_j, f_{j-1})/P(f_j|t_j, f_{j-1})$, we have already divide all of the syllables into two classes: tone-reliable and tone-unknown syllables. That is, we can significantly simplify the probability ratio for tone-reliable case. The probability
density distribution $P(f_j | f_{j-1})$ can be viewed as the weighted sum of five distributions related to the five tones. By assuming that tone-reliable syllables usually have larger values $P(f_j | f_{j-1})$, the five distributions in $P(f_j | f_{j-1})$ are simplified into relatively narrow density distributions with higher probability value. We can therefore approximate the probability ratio $P(f_j | f_{j-1})$/P(t_j | f_{j-1}) by 1/C, where C may be a tone-dependent constant ratio, but we assume it as a tone-independent constant for simplification. For the tone-unknown group on the other hand, $P(f_j | f_{j-1})$ should be a low but highly relative fluctuating value. We thus simplify the problem by assuming it is just a threshold constant probability $P_{th}$. Reformulating Eq. 10 yields the prosodic score approximation

$$P(f_j | t_j, f_{j-1}) = \begin{cases} \frac{1}{C} P(t_j | f_j, f_{j-1}), & \text{tone is reliable given } f_j, f_{j-1} \\ P_{th}, & \text{tone is unknown given } f_j, f_{j-1} \end{cases}$$

(11)

2.5. Prosodic model scores

Taking the log likelihood of Eq. 11 and normalizing the value by $\log(P_{th})$, we define the prosodic model score of character $j$. We reformulate the scoring function of each character $j$:

$$S_{pm}(w_j) = \log \left[ \frac{P(t_j | f_j, f_{j-1})}{C \cdot P_{th}} \right] - \log (P_{th})$$

$$= \begin{cases} \log \left[ \frac{P(t_j | f_j, f_{j-1})}{C \cdot P_{th}} \right], & \text{tone is reliable given } f_j, f_{j-1} \\ 0, & \text{tone is unknown given } f_j, f_{j-1} \end{cases}$$

(12)

The five tone distribution $P(t_j | f_j, f_{j-1})$, a aligned syllable has, is estimated from our prosodic model. We select the probability $P(t_j | f_j, f_{j-1})$ with the tone $t_j$ from the hypothesis character $w_j$, and assign the probability value $P(t_j | f_j, f_{j-1})$ into the scoring function, Eq. 12.

After confirming a syllable as tone-reliable, the model assigns a bonus or a penalty by selecting the probability $P(t_j | f_j, f_{j-1})$ for the tone number $t_j$ of character $w_j$ from the output of the basic tone model. That is, if the probability is larger than the specific value $C \times P_{th}$, the prosodic model adds a bonus for the character $j$. And, if the probability is smaller than that value, the prosodic model gives a penalty. Otherwise, with a tone-unknown pattern, the character gets a zero as its prosodic model score.

Normalizing the score by $\log(P_{th})$ is important because otherwise, negative scores for tone-unknown syllables result in a bias toward arcs with fewer syllables, which leads to an unacceptably large number of character deletions.

2.6. Feature parameters

As input, the prosodic model requires a set of features for each recognized hypothesis syllable, which corresponds to many potential characters. That is, for each set of hypothesis initial-final boundaries in the word graph, we extract a set of prosodic syllable features.

Prosodic information can be derived in 3 dimensions: pitch contour, syllable duration, and energy. For each recognized syllable, various pitch-related features are calculated using the raw pitch contour of the final. For example, the average within-final pitch, the average pitch of several final ending frames, the pitch slope at the beginning frames of the final, and so on. Also, a similar set of pitch-contour-related features are extracted from the pitch contour normalized by its principle components. Each recognized final contour is normalized into a ten-point contour by piecewise linear smoothing. We then map the ten-point contour to the six principle dimensions, the result of principle component analysis performed on the development set. The three top principle components are shown in the Fig. 3. The first component is related to average pitch, the second component describes the rising and falling of the pitch contour, and the third component is related to the low (third) tone. After calculation of the six dimension weights, the normalized contour can be restored by summing the weighted six principle components, after which we extract the set of pitch-related features.

Duration features such as initial and final durations, syllable duration, and the existence of pre-syllable silence are extracted as factors that influence prosody. These features are normalized to the average hypothesis syllable duration of the utterance in question. For energy-related features, we calculate the average log energy for the final and the average energy difference between syllable final and its preceding syllable final. The log energy value of each frame is pre-normalized by the average energy of all hypothesis finals in the utterance.

Finally, we extract 45 prosodic features: 2 sets of 17 pitch features, 7 duration features, and 4 energy features. All features are numerical values, and no categorical features are used in our proposed model. Besides, as three prosodic dimensions are normalized by factors estimated from the hypothesis initial-final alignment for the utterance in question, all the prosodic features are extracted after the end of the utterance.

![Fig. 3. The three top principal components of pitch contour.](image)

3. Experiments

3.1. Corpus and experimental setup

Our corpus was taken from Sinica Continuous Speech Prosody Corpora (COSPRO). The Phonetically-Balanced Speech Corpus (COSPRO 01) is used to train the prosodic model. In COSPRO 01, 2,989 utterances (more than 100,000 syllables) was produced by 2 male and 3 female native speakers. Half of the syllables were used to develop the prosodic model, and the other half were taken as one of the tone recognition testing sets. Also, another 6 speakers, 3 female and 3 male, were selected as testing sets from Multiple-Speaker Speech Corpus (COSPRO 02), in which each speaker recorded 90 to 99 utterances with lengths ranging from 10 to 20 syllables.

The tone classification experiments were done with the help of the corpus alignment marks. The Large-vocabulary speech recognition experiments were performed with a lexicon with 60K entries, a trigram language model, and an intra-syllable right-context-dependent initial/final acoustic model set. The conventional 39-dimension MFCC feature set was used to train the acoustic model.
3.2. Experimental result for tone classification

<table>
<thead>
<tr>
<th>Open test set</th>
<th>Tone accuracy (for tone-reliable syllable)</th>
<th>Tone-reliable percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>COSPRO_01</td>
<td>0.932</td>
<td>0.859</td>
</tr>
<tr>
<td>F001</td>
<td>0.862</td>
<td>0.716</td>
</tr>
<tr>
<td>F002</td>
<td>0.903</td>
<td>0.735</td>
</tr>
<tr>
<td>F003</td>
<td>0.913</td>
<td>0.755</td>
</tr>
<tr>
<td>M001</td>
<td>0.927</td>
<td>0.803</td>
</tr>
<tr>
<td>M002</td>
<td>0.821</td>
<td>0.578</td>
</tr>
<tr>
<td>M003</td>
<td>0.863</td>
<td>0.669</td>
</tr>
</tbody>
</table>

Experiments on speaker-independent tone classification are presented in the Table 1. Here, no speaker normalization was performed on the testing sets. Prosodic features were normalized to the utterances only, or not normalized at all. The “Tone accuracy” column shows the classification accuracy for tone-reliable syllables. The “Tone-reliable percentage” column illustrates what percentages of syllables were classified as tone-reliable. The first row, COSPRO_01, is speaker-dependent set, so it has top tone accuracy. For each other row in Table 1, the result of a speaker for each row, all six speakers are unseen by the prosodic models. Clearly, the table shows that tone recognition accuracy is robust for tone-reliable syllable; as tone patterns become more indistinguishable, the percentage of tone-reliable syllables drops to prevent serious accuracy degradation.

3.3. Experimental results for LVCSR

Four speech recognition tests are presented in the experimental results in Table 2. The baseline method is the conventional speech recognition result with the initial/final acoustic model and the trigram language model. The non-tone prosodic model in [4] is also implemented, which used a different prosodic model not explicitly using linguistic knowledge for character-tone mapping. We report two recognition accuracies, one with all characters in arcs rescored with the recognized tone, and the other, the proposed approach, with a tone-reliable test before arc rescoring.

<table>
<thead>
<tr>
<th>Open test set</th>
<th>Baseline</th>
<th>Non-tone prosodic model</th>
<th>Complete rescoring</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>F001</td>
<td>76.45</td>
<td>77.44</td>
<td>78.17</td>
<td>80.03</td>
</tr>
<tr>
<td>F002</td>
<td>77.57</td>
<td>77.97</td>
<td>81.09</td>
<td>81.32</td>
</tr>
<tr>
<td>F003</td>
<td>77.62</td>
<td>77.76</td>
<td>80.99</td>
<td>82.38</td>
</tr>
<tr>
<td>M001</td>
<td>76.93</td>
<td>76.97</td>
<td>81.40</td>
<td>82.19</td>
</tr>
<tr>
<td>M002</td>
<td>57.43</td>
<td>59.67</td>
<td>60.74</td>
<td>62.66</td>
</tr>
<tr>
<td>M003</td>
<td>76.58</td>
<td>77.06</td>
<td>80.47</td>
<td>81.95</td>
</tr>
<tr>
<td>Average</td>
<td>73.76</td>
<td>74.48</td>
<td>77.14</td>
<td>78.42</td>
</tr>
<tr>
<td>Error reduction</td>
<td>--</td>
<td>2.7%</td>
<td>12.9%</td>
<td>17.8%</td>
</tr>
<tr>
<td>Average excluding M002</td>
<td>77.03</td>
<td>77.44</td>
<td>80.42</td>
<td>81.57</td>
</tr>
</tbody>
</table>

In Table 2, the first six rows are for six speaker, followed by the average. The baseline character (or word) accuracy was about 73.76%, which was raised to 74.48% with the non-tone prosodic model. Then 77.14% character accuracy was achieved by rescoring all characters with their recognized tone type, while the proposed partial rescoring yielded a 78.42% character accuracy. Note that no speaker prosodic feature adaptation was performed in these experiments, and all sets shared the same language model and prosodic model weights in each column. With the help of the proposed prosodic model, a 17.8% error reduction was obtained as compared to the baseline; thus the tone-reliable test allows for a larger prosodic model weight, which yields a 5.6% error reduction over complete rescoring.

The set M002’s poor baseline accuracy may be attributed to the unmatched acoustic model. However, a 5.23% increase in character accuracy was still evident in this set. For the other five sets, the baseline character accuracy was 81.57%, and a 19.8% error reduction was achieved for the five sets.

4. Conclusion

We propose a prosody-incorporated Mandarin speech recognition process, in which tone is considered as an important but not necessary factor in speech recognition. By only focusing on tone-reliable syllables, the prosodic model produces a reliable likelihood score for each syllable. Our experiments show improved performance due to the proposed structure, which can be applied to most existing tone recognizers: a fair test yielded a 17.8% error rate reduction in Mandarin Chinese LVCSR.

5. References