CART-BASED MODELING OF CHINESE TONAL PATTERNS WITH A FUNCTIONAL MODEL TRACING THE FUNDAMENTAL FREQUENCY TRAJECTORIES

Jinfu Ni†‡, Shinsuke Sakai†‡, Tohru Shimizu†‡, and Satoshi Nakamura†‡

†National Institute of Information and Communications Technology, Japan
‡ATR Spoken Language Communication Research Labs, Japan

ABSTRACT

We propose an approach to modeling Chinese tonal patterns, focusing on the basic fundamental frequency ($F_0$) patterns characterized by the contextual linguistic features that can be directly extracted from text. We analyze tonal patterns as sparse target points ($F_0$ peaks and valleys) and represent them in parametric form within the framework of a functional $F_0$ model. The relationships between the target points and underlying linguistic features are trained using classification and regression tree analysis (CARTs), and this functional model is used to trace the $F_0$ trajectories when training the CARTs and to synthesize a tonal pattern from the target points predicted by the CARTs. Our experiments indicate that the proposed method has low $F_0$ prediction errors. Utilization of the $F_0$ ranges measured from training samples could significantly reduce the influences of differences in voice ranges on training a speaker-independent model. Furthermore, the most important roles in characterizing tonal patterns were played by a few linguistic features such as lexical tone context and the distinction between voiced and unvoiced initials.

Index Terms—Prosody modeling, machine learning, functional $F_0$ model, speech synthesis, speech processing

1. INTRODUCTION

Chinese has five lexical tones (Tones 0–4), each characterized by fundamental frequency ($F_0$) contours that coincide with the syllables, thus forming tone patterns. The modulation of the tone patterns can import emphasis to some words and reflect the intonation class of an utterance (statement, question, affirmations, etc.), thus giving intonation patterns. In this paper, we follow from the distinction between neutral and expressive intonation in that neutral intonation reflects language-assigned information which can be characterized by contextual linguistic features and expressive intonation reveals the makeup information that is added by speakers when uttering. We will use term tonal patterns to denote the tone patterns modulated by the neutral intonation. As human listeners make heavy use of the prosodic cues in the understanding process, such cues evidently carry considerable useful information in spoken language systems [1]. However, most systems use prosodic cues in limited, unprincipled ways because there is no established method to employ them.

In speech synthesis, approaches based on hidden Markov models (HMM) have been successfully used to model prosody [2][3]. Furthermore, significant progress has been made in corpus-based speech synthesis technology [4]. These two developments have led to improvement in the quality of synthetic speech, which in turn has led to greater commercialization of this technology [1]. The problem is that for some applications, reading-style speech is no longer adequate because it lacks the aspects of communication conveyed by expressive intonation [5]. We are dealing with the problem by separately modeling the tonal patterns and expressive intonation using the functional $F_0$ model in [6][7]. This paper presents our work on the former (i.e., modeling the tonal patterns); preliminary results on the latter are presented elsewhere [8]. The rest of this paper is organized into three sections: outline of the approach, experimental evaluation, and conclusions.

2. OUTLINE OF THE APPROACH

Figure 1 illustrates the characteristics of the proposed method. Basically, it consists of four components: First, we analyze the tonal patterns in a set of training samples as sets of sparse target points and represent the tonal patterns in parametric form within the framework of the functional $F_0$ model [6] [7]. These target points are then converted to three training parameters, called $m$-parameter, $t$-parameter, and $f$-parameter, which will be defined in Sec. 2.2. Second, we train three CARTs, classification and regression trees (CARTs), called $m$-tree, $t$-tree, and $f$-tree, to model the relationships between respective training parameters and the underlying linguistic features (described in Sec. 2.2). Third, a set of training parameters are predicted by the three trees according to the linguistic features extracted from input text. The predicted training parameters are further converted to a sequence of target points, given the phone boundary information. Finally, these target points are used as model parameters to directly control the functional $F_0$ model and synthesize a tonal pattern for the input text. By using this functional $F_0$ model to trace the observed $F_0$ contours (i.e., the tonal patterns used for training the CARTs), the approach can refine these trees by minimizing the mismatch between the observed and predicted $F_0$ contours.
2.1. Functional $F_0$ model

We use the functional $F_0$ model suggested in [6] [7] to analyze and synthesize the patterns of $F_0$ contours. A model of the $F_0$ patterns basically consists of tonal and intonation components, even in a statistical model [3]. In our model, log $F_0$ contours are mapped into a two-dimensional space in a structure-preserving manner. Furthermore, the tonal and intonation components can separately be represented by an orthogonal base of the two-dimensional space. In consequence, the expression of intonation effects on the tonal patterns can be “removed” when training a model in a statistical sense.

Let $F_0(t)$, as a function of time $t$, indicate an $F_0$ contour from $F_0$ range $[f_0, \text{low } F_0 \text{ boundary}]$ and $f_0$, high $F_0$ boundary]; $A(t)$, the tonal component of $F_0(t)$; and $Z(t)$, the intonation component. The modulation of $A(t)$ through $Z(t)$ to form $F_0(t)$ is expressed as a combination of a resonance mechanism and a frequency transformation as follows [6]:

$$
\ln F_0(t) - \ln f_{0h} = \frac{A(A(t), Z(t)) - A(2, Z(t))}{A(1, Z(t)) - A(2, Z(t))}, \quad (1)
$$

where $A(\lambda, \zeta)$ indicates the resonance mechanism below:

$$
A(\lambda, \zeta) = \frac{1}{\sqrt{(1 - (1 - 2\zeta^2)\lambda)^2 + 4\zeta^2(1 - 2\zeta^2)\lambda}}.
$$

In physics terms, $\lambda$ indicates square the frequency ratios of a forced vibrating system like the vocal cords, and $\zeta$ the damping ratios of the system; $\zeta^2 < 0.5$. Specifically, $f_{0h}$ is forcedly mapped to 2 in this work and $f_{0l}$ to 1 by $\lambda$ in Eq. (1).

When focusing on modeling of the tonal patterns as addressed in this paper, an observed $F_0$ contour is basically represented by the tonal component $A(t)$ parameterized by $\lambda$ as a function of time $t (\geq 0)$, while the intonation component $Z(t)$ can be fixed to a default value $\zeta_0$ [6], namely, $Z(t) = 0.156$ (an empirical value, after this, denoted by $\zeta_0$).

On the other hand, the model follows from an assumption that an $F_0$ contour can be analyzed as a series of target points (tonal $F_0$ peaks and valleys). It is assumed that there are $n$ target points on an $F_0$ contour and they are represented by their tonal component as $(t_i, \lambda_i), i = 1, \ldots, n$, where $t_i$ and $\lambda_i$ indicate the time and the $\lambda$ of the $i$th target point, respectively. The connection from the $i$th target point to the next, denoted by $A_i(t)$, is approximated by a family of exponential functions [7]. $A(t)$ is then expressed as concatenation of all the connections. In mathematical terms,

$$
A(t) = \sum_{i=0}^{n} A_i(t, t_i, \lambda_i, t_{i+1} - t_i, \lambda_{i+1} - \lambda_i, k_i), \quad (2)
$$

where the $i$th target point $(0, \lambda_0)$ is assumed at $t_0 = \lambda_1$ and $t_{n+1} = \infty$. The $i$th connection $A_i(t, t_i, \lambda_i, \Delta t_i, \Delta \lambda_i, k_i) = \bigg\{ \begin{array}{ll}
\lambda_i + \Delta \lambda_i [1 - D(t - t_i, \Delta t_i, k_i)], & \text{for } t_i \leq t < t_{i+1}, \\
0, & \text{otherwise}.
\end{array} \bigg\}
$$

where $D(t, \Delta t, k) = \sum_{j=0}^{k} \frac{c(k) j^j}{j!} e^{-c(k)}, \quad t \geq 0.

Parameter $k_i$ adjusts the configuration of $A_i(t)$ [7]. The $k$-dependent coefficient $c(k)$ is determined by solving the following equation:

$$
\sum_{j=0}^{k} \frac{c(k) j^j}{j!} e^{-c(k)} = 0.1.
$$

The model parameters are summarized as follows.

$n$: The number of target points used for a tonal pattern $(t_i, \lambda_i)$: The $i$th target point, $i = 1, \ldots, n$

$f_0$: A low $F_0$ boundary measured from speech samples

$f_{0h}$: The high boundary of the measured $F_0$ range above $k_i$: $k_i$ is fixed at the synthesis phase for simplicity

$Z(t) = \zeta_0$ (i.e., 0.156) when mapping $f_0$ to (or from) $\lambda$.

Accordingly, the model parameters that need to be trained for representing the tonal patterns are $(t_i, \lambda_i), i = 1, \ldots, n$. 

4254
2.2. Linguistic features and the training parameters

The feature extraction process consists of both linguistic and acoustic aspects. We choose syllable as the basic linguistic unit for the feature extraction because syllables are the carriers of lexical tones. A Chinese syllable consists of three components: initial (21 types), final (36 types), and lexical tone. We consider the following contextual linguistic features and then identify the significant features through experiment.

The linguistic features:
- Current syllabic initial (henceforth, denoted by ci).
- Current syllabic final (cf).
- Current syllabic tone (ct).
- Preceding syllabic initial (pi).
- Preceding syllabic final (pf).
- Preceding syllabic tone (pt).
- Pre-preceding syllabic tone (ppT).
- Succeeding syllabic initial (si).
- Succeeding syllabic final (sf).
- Succeeding syllabic tone (sT).
- Next succeeding syllabic tone (nsT).
- Phrase length in the number of syllables (Len).

We assume that there are m target points associating with a syllable, m ∈ {1, 2, 3}. That is, at most three target points are necessary for a lexical tone [6]. The acoustic features extracted from the speech samples are described as follows.

The acoustic features:
- Order in the m targets associating with a syllable (ith).
- Time of the ith target point (ti).
- F0 of the ith target point in Hz (f0i).
- Voice onset time of the syllable (tv).

If the syllabic initial is a voiced consonant, tv takes the start time of the syllabic initial. Otherwise, tv takes the start time of the syllabic final.
- End time of the syllable (te).
- F0 range [f00, f01] measured from the speech samples.

The F0 ranges are used to reduce the influences of differences in voice ranges on the modeling of the tonal patterns; this is confirmed through experiment. We convert f00, f01 ∈ [f00, f01] to λi ∈ [1, 2] using Eq. (1), i = 1, ..., m. This is done with Z(t) = 0 using an iteration procedure, as described in [6].

Three kinds of training parameters are derived from the measured acoustic features of the training speech samples.

The training parameters:
- m-parameter: number of target points associating with a syllable, m ∈ {1, 2, 3}.
- t-parameter: normalized time ti related to the ith target point; ti = (ti - tv) / (te - tv), i = 1, ..., m.
- f-parameter: mainly using λi converted from f00, i = 1, ..., m. We also use f00 and log f00 as f-parameter to train the corresponding f-tree for comparisons.

3. EXPERIMENTAL EVALUATION

3.1. Speech samples

This experiment used 5489 three- and four-syllabic phrases uttered by two native speakers (a male and a female). The use of these isolated phrases is partly because they cover all the kinds of lexical tone combinations in a balanced manner. These samples are divided into the training and test sets.

- Training set: 5074 isolated phrases (FM-set1)
  - 1319 phrases uttered by a female native (F-set1)
  - 3755 phrases uttered by a male native (M-set1)
- Test set: 415 isolated phrases (FM-set2)
  - 146 phrases uttered by the female native (F-set2)
  - 269 phrases uttered by the male native (M-set2)

The F0 contours were extracted from the speech samples at 5 ms intervals using a tool called TEMPO in STRAIGHT. The potential target points were automatically extracted using an algorithm that minimizes the root mean-square error (RMSE) between the observed and reproduced F0 contours. The extra target points (more than three) were then deleted according to tone types. The F0 ranges [f00, f01] measured from the female and male samples were [122 Hz, 353 Hz] and [74 Hz, 196 Hz], respectively. Figure 1 shows an example of the observed tone patterns (the “+” sequence) with eight target points (the circles in the form (tv, λi)) at the top left corner.

3.2. Experimental method

A CART tool [9] was employed for machine learning. From the results of a preliminary experiment, the initial features for ci, pi, and si were grouped into two classes, voiced and unvoiced, and the final features for cf, pf, and sf were grouped into four classes according to the syllable structures ([onset] nucleus [coda]). Two experiments were then conducted on the speech samples: one to determine which f-parameter and what linguistic features were useful for characterizing the tonal patterns, and the other to evaluate the CART-based modeling of the tone patterns. We trained three models, called the female model, male model, and mixing model, with F-set1, M-set1, and FM-set1, respectively. Both RMSE and correlation criteria were used to evaluate the significance of the linguistic features in characterizing the tonal patterns. Another criterion — the absolute errors between the observed and predicted F0 contours — was used to evaluate the performance of the CART-based models with all the training and test sets.

3.3. Experimental results

An example of tonal patterns predicted with the CART-based models is shown at the top right corner of Fig. 1. The accuracy of prediction m-parameter is 89% for the mixing model. The other main results are shown in Figs. 2 and 3 and Table 1, where boldface type indicates open testing results.
Table 1. Mean errors (and SD) between the observed $F_0$ contours and the $F_0$ contours predicted by the CART-based models with $\lambda$-parameter according to speaker-dependent and mixing models.

<table>
<thead>
<tr>
<th>Model types</th>
<th>Female samples</th>
<th>Male samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Closed)</td>
<td>(Open)</td>
</tr>
<tr>
<td>Female model</td>
<td>15.5 Hz (SD:15.4)</td>
<td>17.0 Hz (SD:15.8)</td>
</tr>
<tr>
<td>Male model</td>
<td>20.1 Hz (SD:16.6)</td>
<td>20.7 Hz (SD:16.1)</td>
</tr>
<tr>
<td>Mixing model</td>
<td>17.3 Hz (SD:14.9)</td>
<td>18.9 Hz (SD:15.1)</td>
</tr>
</tbody>
</table>

4. CONCLUSIONS

This paper presented a CART-based approach within the framework of the functional $F_0$ model to modeling the tonal patterns that can be characterized by the contextual linguistic features. Good results were achieved in terms of $F_0$ prediction errors even when using a speaker-independent model trained by female and male speech samples. Further, the most important roles in characterizing the tonal patterns appeared to be played by the lexical tone context and the distinction between voiced and unvoiced initials in the current and succeeding syllables. Future work will apply these CART-based models to prosody synthesis with the $F0$ modulation technique in [8] by designing an active scale $Z(t)$, as expressed in Eq. (1).

5. REFERENCES