DECISION TREE BASED TONE MODELING FOR CHINESE SPEECH RECOGNITION

Pui-Fung WONG, Man-Hung SIU

Hong Kong University of Science and Technology
Dept. of EEE, Clear Waterbay, Hong Kong
eefung@ust.hk, eemsiu@ust.hk

ABSTRACT

Because of the tonal nature of Chinese languages, correct recognition of lexical tones is necessary for Chinese speech recognition. In order to corporate tone information into Chinese speech recognition, three issues need to be addressed: i) the representation of the syllable pitch contour as well as the tone contour, ii) the lexical tone probability estimation and iii) the integration of tone probabilities into the Viterbi recognition process. In this paper we propose a robust polynomial segmental representation of the pitch contour coupled with a decision tree based tone classifier. We also propose a novel approach of integrating the decision tree tone classifier directly into a single pass recognition process. The proposed approaches were evaluated on tasks of tone classification and tonal-syllable recognition. In regard to tone classification, the robust decision tree gave a tone classification accuracy of 89% for isolated syllables and 71.2% for the continuous speech. Moreover, by incorporating the decision tree tone classifier into the Viterbi search, the tonal-syllable recognition error rate in continuous speech was reduced by 13.54%.

1. INTRODUCTION

It is well-known that Chinese lexical tones depend on the syllable pitch contour. From the linguistic point of view, the shape of a pitch contour determines the tone. For example, the second tone in Mandarin has a rising contour while the fourth tone has a falling contour. Three issues need to be addressed when using tone information for Chinese speech recognition: i) the representation of the syllabic pitch contour, ii) the lexical tone probability estimation and iii) the integration of tone probabilities into the Viterbi recognition process.

Two different approaches are commonly used to extract the syllabic pitch contour information, the frame-based and segment-based approaches. For the frame-based approach [1], the short time $F0$ (Fundamental Frequency) value is calculated for each frame together with its first and second order derivatives. Then, these short-time features is used to build a classifier. For the segment-based method [2], a single pitch contour on the syllabic final or the whole syllabic segment is extracted for each unit. As pitch contour is a supra-segmental feature dependent on the whole segment, the advantage of the segment-based approach is that the shape of the pitch contour can be directly modeled. However, this approach is difficult to integrate into the recognition system.

In this paper we focus on the use of segment-based pitch contour extraction in which each contour is represented as a polynomial function of time, similar to what is done in the polynomial trajectory model in speech recognition [3]. The resulting polynomial coefficients, capturing the shape of the contour, act as tone features. To make the polynomial coefficients insensitive to pitch estimation or boundary errors so that they represent different contour shapes better, we propose the use of robust regression techniques in estimating the polynomial together with using orthogonal polynomial basis. The resulting robust polynomial regression coefficients, together with other tone-related features are modeled using a tone classifier. Most commonly, a neutral network [2, 4] and the HMM [1, 5] were proposed by other researchers for classification tasks. In our case, the decision tree is used for the tone classification because it is widely used in other speech recognition related tasks [6, 7] and has advantages that it is non-parametric, and it can accept both numerical and categorical features. The resulting classifier can be intuitively understood and allowed for the exploration of factors that affect tone production in Mandarin speech. Similar to our work in [8], the tone probabilities generated by the decision tree can be directly integrated into a one-pass recognition algorithm. This addresses one of the major disadvantages of using a segmental-based pitch contour.

This paper is organized as follows. In Section 2 the contour extraction is described. The focus is on how to make the extraction robust to various estimation errors. In Section 3 the decision tree based tone classifier and its integration into recognition are described. Then, the experimental setup and results of isolated tone recognition, continuous tone and syllable recognition, are reported in Section 4. Finally, summary and conclusion are given in Section 5.

2. CONTOUR MODELING

Pitch contour, as well as $F0$ contour, can be represented using a polynomial function in time similar to that used in the segmental model [3]. $F0$ extraction, in our case, involves using CEP [9]. In addition, a median filter and an utterance mean normalization method are used to minimize the doubling/halving errors and inter-speaker variation.

Denote $F = (F_1, F_2, .., F_N)'$ as a sequence of F0 for a given segment and $\hat{F} = (\hat{F_1}, \hat{F_2}, .., \hat{F_N})'$ as an estimated of F. The objective of contour extraction is to find the best polynomial of order $d-1$ with coefficients β_k such that

$$\hat{F_i} = \beta_0 + \beta_1 t_i + \beta_2 {t_i}^2 + \cdots + \beta_{d-1} t_i^{d-1}, \qquad (1)$$

where $t_i = \frac{i}{N}$is a normalized time scale to map the segment duration to 1. Equation 1 can also be expressed in matrix form as,

$$\begin{bmatrix} \hat{F_1} \\ \hat{F_2} \\ \vdots \\ \hat{F_N} \end{bmatrix} = \begin{bmatrix} 1 & t_1 & {t_1}^2 & \dots & {t_1}^{d-1} \\ 1 & t_2 & {t_2}^2 & \dots & {t_2}^{d-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_N & {t_N}^2 & \dots & {t_N}^{d-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{d-1} \end{bmatrix} \qquad (2)$$

$$\hat{F} = T\vec{\beta}$$

0-7803-8484-9/04/$20.00 ©2004 IEEE

The least square error solution for $\vec{\beta}$ is given by

$$\vec{\beta} = (T^{'}T)^{-1}T^{'}F \tag{3}$$

2.1. Robust Regression and Orthogonal Polynomials

The polynomial coefficients estimated using the least square criteria are not robust against outliers. In pitch contour estimation, the sources of outliers can be $F0$ estimation errors (doubling or halving) or syllabic boundary errors. To minimize the impact from outliers, robust regression, which ignores $k\%$ of the worst performing observations [10] during the polynomial coefficients estimation, can be used instead. Let N be the number of $F0$ observations in a segment, the robust estimation procedure is as follows.

```
1  estimate $\hat{\beta}$ using all $N$ pts,
2  compute the square error $E_i$ for each observation $i$,
3  sort $\hat{E}_i$.
4  remove the $kN$ observations with the largest
   square error,
5  and re-estimate $\hat{\beta}$ using $N - kN$ pts.
```

When a polynomial is represented by the set of polynomial basis, i.e. $\{1, t, t^2\}$, the shape of the contour such as the pitch offset, slope and curvature, coupled with the polynomial coefficients. This can make tone classification more difficult. Instead of representing the polynomial as a sum of the powers of t_i, it can be decomposed the polynomial into a set of orthogonal polynomials, $\bar{P}_k(t_i)$'s. Let $f(t_i)$ be a polynomial function with $d-1$ order,

$$\begin{aligned} f(t_i) &= \beta_0 + \beta_1 t_i + \cdots + \beta_{d-1} {t_i}^{d-1} \\ &= \kappa_0 \bar{P}_0(t_i) + \kappa_1 \bar{P}_1(t_i) + \cdots + \kappa_{d-1} \bar{P}_{d-1}(t_i), \end{aligned}$$

where $\{\bar{P}_k(t_i)\}$ is the set of orthogonal polynomial basis, such as

$$\int_0^1 \bar{P}_i(t) \bar{P}_{j \neq i}(t) dt = 0 \; \forall i, j, \tag{4}$$

with $\hat{\kappa}$'s as new polynomial coefficients. With the orthogonal polynomial decomposition, the slope of the tone pattern depends only on the coefficient of $\bar{P}_1$ and the offset of the tone depends only on $\bar{P}_0$.

3. DECISION TREE BASED TONE CLASSIFIER

The decision tree is widely used in speech applications [6, 7] partly because of its non-parametric nature and its ability to handle heterogeneous features. It is composed of a set of leaves nodes and a set of non-terminal nodes, each of which consists of a binary question on the features so as to partition the data into two subtrees. If a tree is used as a classifier, for example, for classifying lexical tones, each terminal node is labeled by the dominant class. If a tree is used to estimate class probability, each terminal node represents a particular class distribution.

Our tone modeling decision tree was trained using the C4.5 decision tree design package [11] utilizing the gain criterion that minimizes class entropy. How the tree is used and the set of features are selected depend on the tasks. In this paper we consider 3 tasks: isolated syllable tone classification, continuous speech tone classification, and tonal syllable recognition.

3.1. Isolated Syllable and Continuous Speech Tone classification

For tone classification, the decision-tree is used as a classifier. Isolated syllable tone recognition is a simpler problem because there are only four different tone patterns and most are fairly stable. Apart from the orthogonal polynomial coefficients, other prosodic features including syllable duration, syllable energy and regression error can be included as tone features. Moreover, each tone pattern in Mandarin Chinese has a different duration distribution. For example, on average, tone 3 is the longest and tone 4 is the shortest. If the energy level is high, $F0$ estimation is more reliable as $F0$ estimation error is sensitive to the frame energy level. In addition, intuitively, the regression coefficients of segments with large regression errors are less reliable.

Continuous speech tone recognition is more difficult than isolated syllable because of the addition of a neutral tone (tone 0), context-dependent effects, and co-articulation effects. Recognizing tone 0 is very difficult because it is usually highly context dependent and its $F0$ contour pattern is relatively arbitrary. Furthermore, the instantiation of a lexical tone in continuous speech depends on the neighboring syllables, commonly known as tone sandhi [12]. Their occurrences also depend on the word segmentation and phrase breaks. Moreover, because of co-articulation effect, the contour shape of a syllable may be affected by the $F0$ contour patterns of neighboring syllables. In order to cope with these effects, the tone features are expanded to include contextual features as shown in Table 1.

- Orthogonal polynomial coefficients, log-energy, regression error, durations of preceding and succeeding segments
- duration and log-energy of both unvoiced or silence segments before and after the current syllable,
- the type of initial such as unvoiced initial, null initial, voiced initial, before and after the processing final, and
- two binary indicators to specify whether the current syllable is the first or last syllable of the sentence

Table 1. Context dependent features for continuous speech tone recognition.

3.2. Integration of tone information into recognition

Tone probabilities can be integrated into the recognition search process as suggested in [8]. The idea is that the tone likelihood of each possible syllables are computed and incorporated at the syllable ends. In this case, the decision tree serves as a probability estimator. However, because the decision tree computes the tone posterior probability instead of the likelihood, it has to be converted into a likelihood (similar to what is done in neural networks) before it can be combined with the acoustic likelihood in HMM. Define $\mathcal{D}(\vec{\Gamma_j})$ to the corresponding terminal node of the decision tree given tone feature $\vec{\Gamma_j}$.

The tone posterior probability $p(\tau_j = i|\vec{\Gamma_j}) = p(\tau_j = i|\mathcal{D}(\vec{\Gamma_j}))$. Using Bayes' rule, the tone likelihood is given by,

$$\begin{aligned} p(\vec{\Gamma_j}|\tau_j = i) &= \frac{p(\tau_j = i|\vec{\Gamma_j})p(\vec{\Gamma_j})}{p(\tau_j = i)} && (5) \\ &\propto p(\tau_j = i|\mathcal{D}(\vec{\Gamma_j})), && (6) \end{aligned}$$

Then the term $p(\vec{\Gamma_j})$ is the prior probability of tone feature which is not a function of the tones and can be computed during training.

3.2.1. Searching process

To recognize Chinese syllables or words, the tone likelihood, which is an extra term, is computed at the end of each syllable. In order to evaluate the tone likelihood, syllable segmentation is needed. Using the token passing approach [8] for recognition, the best syllable begin time is stored in the partial path alignment and is available at any time instance. At time t, suppose j is the end node of syllable k which has the tone of τ_j. Denote $\mathcal{B}(t, j)$ the best word begin time associated with end node j at time t. The tone feature, denoted $\Gamma(\mathcal{B}(t, j), t)$ can be computed using polynomial regression. Integrating the tone likelihood into the word end score, we have

$$\psi_t(j) = \psi_t(i) + \log(a_{ij}) + \log p(\Gamma(\mathcal{B}(t, j), t)), \qquad (7)$$

where $\psi_t(j)$ is the Viterbi score at time t for state j. It is interesting to note that the tone likelihood can be thought of as a sort of word insertion penalty.

4. EXPERIMENTS

Experiments were performed on the tasks of tone classification and tonal syllable recognition. Our recognition system is HMM-based and uses a left-to-right topology without skips. Our baseline system used 39 features, consisting of 12 MFCC plus frame energy, and their first and second order derivatives. In both cases, the acoustic units were context independent syllable initials and finals. These are commonly used as the basic acoustic units in Mandarin speech recognition. The acoustic inventory included 24 initials [13] (included null initials) and 165 tonal finals plus two extra units for silence and short pauses.

4.1. Isolated Syllable Tone Classification

The first set of experiments performed were on isolated tone classification. Our goal was to derive a suitable configuration for pitch contour modeling and to develop a basic context-independent decision tree tone classification system. We evaluated three different conditions: i) the best polynomial representation, such as the order of polynomial and the basis functions, ii) the effectiveness of the robust estimation and, iii) the optimal set of decision tree features.

For these experiments, the HKU93 corpus [14], which consists of a total of 20 native speakers, 10 females and 10 males, was used. The corpus is designed to cover all the tonal syllables. So, each speaker pronounces all syllables in all tones at least once. Speech data from 14 speakers, 7 males and 7 female, were used for training, and the remaining data were used for testing. There are in total 23342 and 9395 sentences in the training and test set respectively.

Table 2 shows the relative gain of using orthogonal polynomials with different polynomial orders. In there experiments, decisions were built using only the polynomial coefficients as features. The classification errors indicat that the recognizer has the minimum classification error at $d = 5$ (a fourth order polynomial) and that the classification error is reduced by 2.6% from 18.89% to 15.25% if the pitch contour is represented by a set of orthogonal polynomials.

Dimension of feature vector (d)	Tone Classification Error (%)		
	Original Polynomial	Orthogonal Polynomial	Relative Gain
3	34.70	33.32	1.38
4	21.14	18.56	2.58
5	18.89	15.25	2.64
6	28.21	20.66	7.55

Table 2. Tone classification on Mandarin Chinese isolated syllable varied different dimension d of feature vector and orthogonal transformation.

Errors in F0 estimation and syllable boundary estimation can affect the regression coefficients estimation for the pitch contour. We performed a set of experiment so as to investigate the effectiveness of robust regression against these errors using fourth order orthogonal polynomials to represent the tone contours. The tone classification results, with different percentages of observations excluded in the robust regression, were tabulated in Table 3.

The results show that with a limited percentage of observations excluded from the regression coefficient estimation, accuracy improves. However, when too many are excluded, the reduction in data out-weighed the gain from increased robustness. Compared to using traditional regression in which all data are used, the robust regression with 20% of the observations ignored results the classification error from 16.5% to 11.6%. Apart from polynomial coefficients, we also investigated the addition of other features such as regression error, syllable duration and syllable energy. By adding these extra features, the tone classification error rate overall was further reduced from 11.6% to 10.3%, another 10% relative improvement.

Percentage of data Ignored (k%)	0	10	20	30	40
Tone Classification Error(%)	16.5	12.8	11.6	12.4	12.7

Table 3. Experimental results on tone classification by using robust regression.

4.2. Tone classification and Tonal Syllable Recognition on Continuous speech

Based on what we had learned from the isolated syllable experiments, we shifted our attention to continuous speech. A different corpus, the Chinese 1998 National Performance Assessment (Project 863) [15] was used. 863 is a Mandarin corpus of continuous speech with sentences taken from the People's Daily newspaper between 1993 and 1994. 863 contains only 2,443 unique sentences as each utterance is read multiple times and by multiple speakers. In our experiments, only speakers with Beijing accents were selected. 34 speakers (17 male and 17 female) were selected for training and another 12 for testing. To ensure that the test was fair, the test set was designed such that there was no overlap between training and test speakers, nor any overlap in sentences.

Tone classification and syllable recognition in continuous speech is much more challenging than that in isolated syllables. In addition to typical continuous speech issues such as variation in speaking rate, style, emotion, the high level of the co-articulation effect [4] between syllables in Chinese, and the presence of tone sandhi [12] which modifies the lexical tone of a word based on the lexical tones of the neighboring syllable makes processing continuous Chinese more challenging. In addition, an extra neutral tone, used only in continuous speech [12] in filler words or to change emphasis, was presented.

We performed three sets of experiments. In the first set, we used a configured system similar to that in isolated syllable tone classification. For tone classification on continuous speech, this system did not use any context information. As contextual information provides important information in continuous speech, we experimented in our second set of experiments with a context dependent tone classifier for continuous speech. In our third set of experiments, because the HMM models are context independent, we incorporated the context-independent tone classification decision tree into the Viterbi recognition process.

In the first experiment, the construction of the decision tree for continuous speech was very similar to that did on an isolated word with the exception that a third order polynomial was used. Furthermore, the training data was re-sampled between the tone classes in order to minimize the effect of an unbalanced prior between the phone classes during the construction of the decision tree. The tone classification results are tabulated in the first row of Table 4.

To capture some of the contextual effects that are quite significant in Mandarin, we expanded the features used in decision tree to include information from the left and right neighboring syllables as summarized in Table 1. The results, after adding left-context, right-context and both contexts, are tabulated in Table 4. Notice that the tone classification error is reduced by 3.6% (10% relative) with the use of context dependent tone features, and that the classification results of all tones improved. In particular, the classification error of tone 0 dropped significantly from 61.8% to 31.0%. This is consistent with our understanding that the tone pattern of tone 0 is highly dependent on its neighboring tone. Moreover, the result of using left context only is better than that of the right for tones 0 and 2, suggesting that the tone of the previous syllable has more impact than that of the next syllable.

In our third set of experiments, we integrated the decision tree into the Viterbi recognition process. The tone classification decision tree was constructed using context-independent tone features as the HMM models were also context independent. Because of our focus on the acoustics, no language model information was used in this experiment. The recognition results with, and without the decision tree tone probability are summarized in Table 5, which the tonal-syllable error rate dropped from 52.52% to 42.16%.

Experiment	Tone Classification Error%					
	Recognized Tone					Total
	0	1	2	3	4	
CI	61.8	18.7	25.4	48.4	31.6	31.8
left CD	51.3	21.0	33.0	46.4	31.4	30.7
right CD	55.1	19.9	25.6	46.8	32.1	31.6
left+right CD	31.0	18.3	24.8	40.6	30.3	28.2

Table 4. Tone classification on continuous speech, where CI is context-independent tone features and CD is context-dependent tone features.

Experiment	Tonal-syllable% Error Rate%	Relative Gain
MFCC only (baseline)	52.52%	-
C-I decision tree	45.41%	13.54%

Table 5. Results of tonal syllable recognition by integration of the decision tree which is constructed by context-independent tone features.

5. CONCLUSION

In this paper we proposed a robust segment-based pitch contour representation together with a decision tree based classification system under the decision tree framework. In tone classification, using robust features and orthogonal polynomial representation together with context dependent tone related features, we obtained a tone error rate of 11% tone in isolated syllable tone classification. In addition, the tone error was reduced from 32% to 28% in continuous speech after including context-dependent features. Using our proposed integration approach that can directly evaluate the tone probability of a possible syllable within the recognition process, the tonal-syllable recognition error was reduced by close to 14%.

6. ACKNOWLEDGMENTS

This work is partially supported by the Hong Kong RGC under the grant under HKUST 6049/00E. The authors would like to acknowledge Wilson Tam for sharing his recognizer with us.

7. REFERENCES

[1] C.H. Huang and F. Seide, "Pitch tracking and tone features for mandarin speech recognition," in *Proceedings of ICASSP*, 2000, pp. 1523–1526.

[2] P.C. Chang, S.W. Sun, and S.H. Chen, "Mandarin tone recognition by multi-layer perceptron," in *Proceedings of ICASSP*, 1990, pp. 517—520.

[3] H. Gish and K. Ng, "Parametric trajectory models for speech recognition," in *Proceedings of ICSLP*, 1996, pp. 466–469.

[4] S.H. Chen and Y.R. Wang, "Tone recognition of continuous mandarin speech based on neutral networks," *IEEE Transaction on Speech and Audio Processing*, vol. 3, pp. 146–150, Mar 1995.

[5] W.J. Yang, Y.C. Lee, Y.C. Chang, and H.C. Wang, "Hidden markov model for mandarin lexical tone recognition," in *IEEE Transaction on ASSP*, Jul 1998, vol. 36.

[6] G. Boulianne and P. Kenny, "Optimal tying of hmm mixture densities using decision trees," in *Proceedings of ICASSP*, 1996, pp. 350–353.

[7] L.R. Bahl, P.F. Brown, P.V. de Souza, and R.L. Mercer, "A tree-based statistical language model for natural language speech recognition," *IEEE Transaction on ASSP*, vol. 37, pp. 1001–1008, Jul 1989.

[8] P.F. Wong and M.H. Siu, "Integration of tone-related features for mandarin speech recognition by a one-pass search algorithm," in *Proceedings of ISCSLP*, 2002, pp. 64–68.

[9] P.F. Wong and M.H. Siu, "Integration of tone-related feature for chinese speech recognition," in *Proceedings on ICMI*, 2002, pp. 476–479.

[10] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*, John Wiley and Sons, New York, 1987.

[11] J.R. Quinlan, *C4.5 Programs for Machine Learning*, Morgan Kaufmann series in machine, 1993.

[12] L.S. Lee, C.Y. Tseng, and M.O. Yong, "The synthesis rules in a chinese text-to-speech system," *IEEE Transaction on ASSP*, vol. 37, pp. 1309–1320, Sep 1989.

[13] S. Gao, T. Lee, Y.W. Wong, B. Xu, P.C. Ching, and T.Y. Huang, "Acoustic modeling for chinese speech recognition: a comparative study of mandarin and cantonese," in *Proceedings of ICASSP*, 2000, pp. 1261–1264.

[14] Y.Q. Zu, X.X. Li, and C. Chan, "Hku93 - a putonghua corpus," in *Department of Computer Science, University of Hong Kong*, 1994.

[15] R.H. Wang, "National performance assessment of speech recognition systems for chinese," in *Proceedings of Oriental COCOSDA workshop*, 1999, pp. 41–44.