interspeech 2008 incorporating SST 08

Mandarin Chinese Tone Nucleus Detection with Landmarks

Siwei Wang, Gina-Anne Levow

Department of Computer Science, University of Chicago, Chicago, IL U.S.A

siweiw@cs.uchicago.edu, levow@cs.uchicago.edu

Abstract

This paper discusses a new approach to improve tone recognition by modeling the tone nucleus with vowel landmark detection. The tone nucleus region is identified based on vowel landmark frames derived by an automatic landmark recognition system. In the corresponding tone recognition experiments, the best results with landmark-based tone nucleus regions outperform the best baseline system results by more than 6%. Moreover, in an exploratory experiment, the tone recognition accuracy using tone nucleus regions based only on vowel landmark evidence shows less than 2% degradation relative to the accuracy obtained using both landmark frames and force-aligned vowel boundary information. These findings further demonstrate the potential to perform tone recognition based on landmark detection alone, without full speech recognition or aligned transcriptions.

Index Terms: prosody, tone recognition, vowel landmark detection, tone segmentation, sonority profile.

1. Introduction

In Mandarin Chinese, there are four canonical tones (High, Rise, Low, and Fall) and one neutral tone. Those tones determine the identity of the corresponding syllable. Tone recognition is necessary for automatic speech processing of Mandarin Chinese and other tonal languages. However, the coarticulation effect from adjacent tones can substantially deform these underlying pitch patterns in continuous speech. Since tones are identified based on their pitch patterns, these tonal coaticulatory effects increase the difficulty of automatic tone recognition.

The hypothetical region containing the best-articulated features corresponding to the syllable's canonical tone target is referred to as the tone nucleus. In past research [2], the location of the tone nucleus was estimated by segmental K-means over the fundamental frequency contour, improving tone recognition accuracy.

Rather than using statistical modeling of the pitch contour to extract the tone nucleus as in [2], we look into landmark detection as described in [1]. In a landmark-based speech recognition system, landmarks are frames corresponding to maxima, minima, and inflection points of particular acoustic events. In [1], vowel landmark detection identifies the frames with the highest sonority during vowel production based on support vector machine learning from MFCC parameters. Since tone is associated with vowels in Mandarin Chinese, we aim to locate tone nuclei around these vowel landmark frames by selecting continuous frames of non-zero pitch values around the landmark with sonority scores above a predetermined threshold

This paper is organized as follows. In Section 2, we introduce the dataset and the feature combination we will employ in all of our experiments. Section 3 introduces two landmark detectors using different parameters to identify tone nucleus regions. These two landmark detectors form the basis for four different settings of landmark-based experiments. To contrast, we employ three baseline models. In the discussion of our results in Section 4, the best of our landmark-based approaches outperforms the best baseline model by more than 6% absolute. More interestingly, the tone recognition based tone nucleus regions identified based only on landmarks shows less than 2% degradation relative to the best experiment which exploits both landmarks and force-aligned vowel boundary information. In further analysis, we found that landmarks occur late in the vowel and yield nuclei focusing on the middle of the vowel. We conclude our work in Section 5.

2. Dataset

2.1. Data preparation

All of our tone recognition experiments are evaluated on a dataset containing 500 utterances with 4800 syllables in total. These utterances are chosen randomly from the Topic Detection and Tracking (TDT2) Voice of America Mandarin Broadcast News corpus from 1998, distributed by the Linguistic Data Consortium [1]. The audio was force aligned to the corresponding automatically word-segmented anchor scripts using the University of Colorado's Sonic Speech Recognizer [5].

2.2. Acoustic features

For tone recognition, the pitch and amplitude contour are extracted using Praat's[3] commands, "To pitch..." and "To amplitude...". Our feature combination considers both nucleus features and contextual features. Nucleus features describe the current syllable, while the contextual features compensate for effects of coarticulation with adjacent syllables through two kinds of features: "extended" features from adjacent syllables and "difference" features computed between the current syllable and its previous and following syllable. Those contextual features have been assessed in [7] and led to improvement of overall accuracy for tone and pitch accent prediction. The details of our feature combination appear below:

- Nucleus features

  Pitch features: mean, standard deviation, max; values at 0, 0.25, 0.5, 0.75, 1 of the tone region; Slope of linear interpolation of pitch contour against time; first and last values of the tone region.

  Amplitude features: mean and max

- Contextual features

  Extended features: slope, mean, max from pitch and amplitude feature of both previous and following

This work was supported by NSF IIS: #0414919

[1] http://www.ldc.upenn.edu

tone segments. Last pitch value of the previous tone segment and the first pitch value of the following tone region

Difference features:(current tone segment against both previous and following tone segments): mean, max, slope of pitch and max of amplitude.

2.3. Baseline Experiment

To contrast with the landmark-based tone nucleus segmentation, we implement 3 baseline models:

1. *whole*: features extracted from the whole vowel region;
2. *half*: features extracted from midpoint to the end of the vowel region. This approach is inspired by the claim [6] that pitch targets are better approximated later in the syllable.
3. *60ms*: features extracted from 60ms window around the landmark position.

3. Experiment

3.1. Landmark speech recognition

Landmark speech recognition described in [1] aims to model pure speech recognition, which decodes the speech waveform into a series of phonological units without invoking higher-level linguistic knowledge. Therefore, this approach highlights a frame/point in speech of particular significance when some distinctive features are observed. Thus it transforms the speech waveform into a sparse point process containing points in time indicating important events or *landmarks* corresponding to maxima, minima, and inflection points of specialized acoustic properties. In the hierarchy of those distinctive features, the root sonorant-obstruent feature distinguishes the sonority profile of the signal from the obstruent. As the most open, full throated sonorant class of sounds, vowels and the firing of vowel landmarks correspond to peaks of the sonority profile and provide anchor points that define syllable-sized analysis units.

3.2. Vowel landmark detection

The sonority profile determining vowel landmarks is quantified by scores from a support vector machine using MFCC parameters, extracted from the corresponding frames. The step between adjacent frames denotes whether the following MFCC parameters will consider part of the current frames or not. The score ranges from -3 to 3. Only the frames with positive scores can be associated with landmarks. The higher the score, the more likely the landmark is correct. In addition, we define $landmarkscore$ to be the locally maximum sonority score corresponding to the landmark frame.

We employ two landmark detectors with different parameters to assess different settings for identifying tone nuclei. The landmark detectors' parameters and their detection rate for our dataset are included in Table 1.

	step length	window size	detection rate
cl_35	20ms	35ms	94.40%
cl_10	10ms	10ms	94.75%

Table 1: Landmark detectors

Using the dataset of 500 utterances as described in Section 2.1, we collect those syllables with landmarks from both landmark detectors to evaluate the tone nucleus modeling in the following sections.

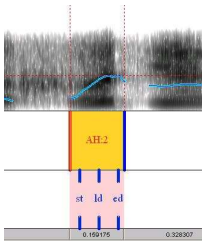

Figure 1: Second tone example: Top: spectrogram with pitch contour, Mid: vowel alignment; Bottom: landmark boundaries and landmark position: st:start, ed:end, ld: landmark position

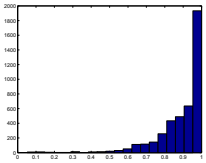

Figure 2: Histogram of the landmark_ratio with threshold: $Landmarkscore - 1$

3.3. An example of landmark-based Tone nuclei modeling

Figure 1 shows an example of the second (rising) tone and its corresponding tone nucleus region generated from landmarks. By thresholding the sonority score with a predetermined value, we generate the landmark-based tone nucleus region where the pitch contour best approximates the canonical second tone rising pattern and excludes the final level region.

3.4. Tone nucleus modeling with landmarks

We employ four different settings for modeling tone nucleus with landmark frames. We distinguish two main categories based on whether or not evidence from vowel boundaries, identified by force-aligned transcription [5], is used in conjunction with evidence from landmark frames to determine the tone nucleus region. We perform five-fold cross validation tone recognition experiments based on tone nucleus regions from those four landmark-based settings and three baseline models. We test all models with Support Vector Machines using RBF kernels[4].

3.4.1. Tone Nucleus Modeling within Vowel Boundaries

The goal of these experiments is to determine if vowel landmark evidence can refine the vowel-based region for tone recognition to isolate the better-articulated tone nucleus. Hence tone nucleus regions are formed with continuous frames inside vowel boundaries of non-zero pitch values around landmark. The sonority score threshold aims to select the best tone nucleus frames inside the vowel. Here we define $landmarkratio$ to be the number of frames with sonority scores above a specified threshold inside a vowel divided by the total number of frames in that vowel.

Based on the histogram in Figure 2, most of the frames inside vowels in our dataset achieve a sonority score no less than $Landmarkscore - 1$. Therefore, we employ the $Landmarkscore - 1$ as the threshold to form the tone nucleus region.

Hence, there are two settings for identifying landmark-based tone nuclei employing two landmark detectors

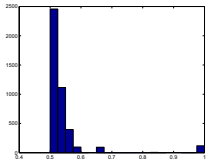

Figure 3: Histogram of landmark_ratio with threshold: $0.7 * Landmarkscore$ For most vowels in our dataset, this threshold extracts at least 50% of the vowel region

(*cl_10_in,cl_35_in*) within vowel boundaries. All of the tone recognition experiments included in this section are evaluated on syllables which fire landmarks in both detectors and which have tone nucleus regions that are longer than 40ms for feature extraction. Table 2 gives the tone distribution of the dataset.

1st	2nd	3rd	4th
906	957	457	1255

Table 2: Tone distribution for experiments in section 3.4.1

From Table 3, we can conclude that both of our landmark-based tone nucleus recognition approaches achieve significantly better results than the best of our baseline models. The best (*cl_35_in*) outperforms the best baseline model by 6% absolute.

3.4.2. Tone Nucleus Modeling without Vowel Boundaries

A landmark is a frame denoting the position of the peak of the vowel sonority profile. In landmark detection, the score from a support vector machine based on MFCC parameters ranks each frame's likelihood of being in a vowel. Hence a fraction of the peak landmark score can serve as a threshold for tone nucleus identification within the vowel.

Figure 3 shows that at least 50% of the frames in all vowels are selected as nuclear using a sonority score threshold of 70% of the landmark peak score. We take this threshold as a conservative approach in our preliminary exploration of tone nucleus segmentation in the absence of explicit vowel boundaries, identifying well-articulated regions of at least 40ms duration.

As in the previous section, we employ two different parametrizations of the vowel landmark detector, but here we use only the landmark-based tone nucleus selection procedure above, with no other vowel alignment information, resulting in two experimental configurations: *cl_10* and *cl_35*. This conservation thresholding method identifies tone nuclei in a subset of the syllables, with the distribution shown in Table 4. We compare the effectiveness of tone recognition employing tone nucleus selection without forced alignment-based vowel boundary information to tone nucleus selection with vowel boundary information, as well as our baseline configurations, on this subset of syllables, shown in Table 5.

1st	2nd	3rd	4th
740	739	379	1002

Table 4: Tone distribution for experiments in Section 3.4.2

The average tone recognition accuracy of the best of the landmark-based tone nucleus identification approaches is 65.31%, obtained in configuration (cl_10_in). This configuration employs the tone nucleus regions identified based on the cl_10 detector within the force-aligned vowel boundaries. It is 4.5% better than the best mean accuracy of 60.7% using the "half" baseline model. Both experiments on tone nucleus regions based on landmarks alone show 2% decrease in accuracy relative to the comparable experiment with known vowel boundaries, but the differences are not significant by Wilcoxon Rank Sum Test. In addition, all tone recognition experiments evaluated on landmark-based tone nucleus regions achieve better results than the three baseline models.

	begin_diff=0	begin_diff<0
end_diff=0	4.7%	18.78%
end_diff<0	15.08%	61.47%

Table 6: Four different positions of tone nuclei and their frequency in the dataset

4. Discussion: Understanding Tone Nuclei

In this section, we aim to better characterize the landmark-based tone nuclei and the relationship between the tone nucleus boundaries and vowel boundaries. Through this characterization, we hope to understand the significant improvements obtained through tone nucleus modeling for tone recognition. To facilitate our analysis, we define the following two measures.

1. begin_diff: $\frac{vst-ldst}{ved-vst}$ We use *ldst* and *lded* to denote the beginning and ending of the landmark-based tone nucleus region, and *vst* and *ved* to denote the beginning and ending of the vowel alignment. We compute the ratio of the difference between the starting position of the landmark-based tone nucleus region and the starting boundary of the vowel to the alignment-based vowel length. This value is negative if the landmark-based nucleus starts in the middle of a tone.
2. end_diff: $\frac{lded-ved}{ved-vst}$ We compute the ratio of the difference between the ending position of the landmark-based tone nucleus region and the ending boundary of the vowel to the alignment-based vowel length. This value would be negative if the landmark-based nucleus ends before the vowel ends.

4.1. Tone Nuclei in Experiments with Vowel Boundary Information

The 6% improvement we obtained from *cl_35_in* relative to the best baseline demonstrates that landmarks are helpful in locating the best-articulated tone production inside a vowel. We first generate statistics with (begin_diff,end_diff) as shown in Table 6. We observe that the majority of tone nuclei exclude frames from both the beginning and end of the vowel region. Then we generate four plots in Figure 4 to further illustrate the position of tone nuclei regions relative to vowel boundaries: It is obvious from the UL plot that most of the landmarks are in the later part of the vowel. This asymmetry is consistent with the pitch target approximation hypothesis of [6] that argues for better articulation of tonal targets later in the syllable and alignment of segmental and tonal articulation. In addition, most of the tone nucleus regions tend to cover more than half of the vowel region, but less than 5% cover all frames inside vowel boundaries. It is also straightforward to observe that the scatterplots concentrate

	1	2	3	4	5	mean
cl_35_in	64.59%	64.73%	68.33%	65.96%	63.76%	65.47%
cl_10_in	64.17%	63.49%	67.66%	65.01%	64.45%	64.96%
whole	57.95%	58.78%	60.03%	58.92%	59.47%	59.03%
half	58.23%	56.71%	61.41%	60.72%	60.17%	59.45%
60ms	46.47%	49.38%	53.80%	52.84%	51.03%	50.70%

Table 3: Contrastive experiment on tone nucleus modeling with landmark exploiting force aligned vowel boundaries: cl_35_in outperforms the best baseline $half$ by more than 6%

	1	2	3	4	5	mean
cl_35	62.41%	61.01%	65.91%	62.41%	64.9%	63.33%
cl_10	62.06%	61.54%	64.5%	64.97%	62.24%	63.06%
cl_35_in	62.41%	63.99%	68.88%	64.86%	66.43%	65.31%
cl_10_in	63.46%	63.46%	67.66%	65.38%	66.61%	65.31%
whole	57.52%	60.66%	63.81%	58.92%	60.31%	60.18%
half	59.62%	56.99%	63.29%	61.36%	62.24%	60.7%
60ms window	48.6%	51.39%	54.72%	53.85%	53.50%	52.41%

Table 5: Contrastive experiment on tone nucleus region generated based on landmark without considering force-aligned vowel boundaries. The result in cl_35 is comparable to cl_35_in with less than 2% degradation.

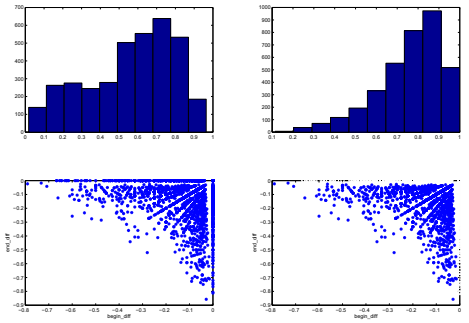

Figure 4: Histograms and scatterplots showing the distribution of tone nucleus ithin vowel boundaries. Upper left (UL) plot: the position of landmark frames inside the vowel; Upper right (UR) plot: the fraction of vowel being covered by the tone nucleus;Lower left (LL) plot: the scatterplot of (begin_diff, end_diff) for the whole dataset;Lower right (LR) plot: the scatterplot of (begin_diff, end_diff) for the vowels with tone nucleus regions that are not adjacent to any vowel boundaries.

around the origin point. It appears that tone nucleus selection excludes frames at both the beginning and end of the vowel similarly. This restriction minimizes effects of both carryover and anticipatory coarticulation.

5. Conclusions and Future Work

To summarize, we demonstrate that the position of tone nucleus can be successfully approximated with its landmark frame and those frames around it with sonority scores higher than a predetermined threshold. Moreover, such an estimation provides consistent segmentation and thus extracts out the best-articulated tonal region from continuous speech.

An immediate extension could be applying this landmark-based tone nucleus modeling to full-profile tone recognition on either Mandarin Chinese or other tonal languages. Furthermore, the results we have obtained from the tone-nucleus regions based on landmark information only inspires possible extension to performing tone recognition without requiring full speech recognition or aligned transcription.

6. References

[1] Aren Jansen, Partha Niyogi, “A Probabilistic Speech Recognition Framework Based on the Temporal Dynamics of Distinctive Feature Landmark Detectors”, Technical Reports, TR-2007-07 “http://www.cs.uchicago.edu/files/tr_authentic/TR-2007-07.pdf”.

[2] Jinsong Zhang, Keikichi Hirose, “Tone nucleus modeling for Chinese lexical tone recognition”, Unveristy of Tokyo, Tokyo, Japan Speech Communication 42 (2004) 447-466.

[3] P. Boersma, D. Weenink, Praat:doing phonetics by computer Version 4.3.01 (2005) Retrieved from “http://www.praat.org”

[4] Chih-Chung Chang, Chih-Jen Lin, “LIBSVM a library for support vector machines” “http://www.cise.ntu.edu.tw/cjlin/libsvm”

[5] B.Pellom, W. Ward, J. Hansen, K. Hacioglu, J. Zhang, X. Yu and S. Pradhn “University of Colorado dialog systems for travel and navigation” (2001).

[6] C.X.Xu, Y.Xu, L.-S.Luo, “A pitch target approximation model for f0 contour in Mandarin” Proceedings of the 14th International congress of Phonetic Sciences, (1999) 2359-2362.

[7] Gina-Anne Levow, “Context in Multi-lingual Tone and Pitch Accent Prediction”, Proceedings of Interspeech 2005, p. 1809-1812.