Improving Tone Recognition with Combined Frequency and Amplitude Modelling

Siwei Wang, Gina-Anne Levow

1100 E. 58th Street
Department of Computer Science
University of Chicago, Chicago, IL,USA
siweiw,levow@cs.uchicago.edu

Abstract

To improve tone recognition in continuous speech, we propose a strategy focusing on separating regions influenced by tonal coarticulation from regions that more closely approximate canonical tone production. Given a syllable segmentation, this approach employs amplitude and pitch information to generate an improved sub-syllable segmentation and feature representation. This sub-syllable segmentation is derived from the convex hull of the amplitude-pitch plot. Our approach achieves a 15% improvement using our segmentation strategy over a simple time-only segmentation. Finally, a future extension with sequential labelling is discussed.

Index Terms: Tone recognition, prosody, amplitude, graphical framework.

1. Introduction

Tone recognition is important in speech recognition for tonal languages like Mandarin Chinese. The tonal system of Mandarin Chinese has four lexical tones (High(1), Rising(2), Low Falling-rising(3) and Falling(4))and one neutral tone. State-of-the-art tone modelling research has demonstrated that the observed fundamental frequency contour in continuous speech often diverges significantly from the underlying canonical tone contour. This deformation is caused by coarticulation, which results from physiological limitations such as maximum speed of pitch change that force smooth transitions between adjacent tones with different pitch heights in continuous speech. Both anticipatory and carryover coarticulation with adjacent tone objects is frequently reported. In addition, tone realization is also affected by broader influences such as phrase and topical change. These variations create significant challenges for tone recognition.

Current models like the parallel encoding and target approximation(PENTA) model [1] and StemML[2] have proposed methods to explain the coarticulatory influence on the surface realization of the underlying canonical tone. For instance, PENTA hypothesizes that the carryover coarticulation dominates tone realization and thus that the true tone is more closely approximated in the latter half of the syllable. Sun[3] used the pitch at the midpoint of the syllable and fit the pitch contour from the midpoint to the end of the syllable for pitch accent recognition, effectively a temporal segmentation. Subsequently, Zhang and Hirose[4] proposed a model which successfully identifies tone “nucleus” regions for canonical tone production. The tone region is segmented by k-means clustering of pitch contour units; the nucleus itself is identified based on features including segmental time and energy. Coarticulation effects and wider context modelling are also emphasized in many current tone recognition approaches [5, 6].

Naturally, most approaches to tone recognition and perception have emphasized pitch information. However, some recent research has identified the importance of amplitude for tone perception as well. In human perception experiments using devoiced tone samples with amplitude only, the tones with declining pitch variation can be correctly recognized[7]. Also, the contour tones with both rising and falling variations have been shown to be perceivable by their amplitude patterns only[8]. Furthermore, cochlear implant experiments successfully enable their native Mandarin-speaking subjects to perceive all 4 tones using amplitude in a sufficient number of bands[9].

From the above human perception experiments, we conclude that amplitude carries important information for tone recognition. In this paper, we describe an approach employing amplitude in segmenting tone regions for better tone recognition. This approach involves a two-step procedure using the amplitude and pitch contour in generating a proper segmentation and a subsequent feature representation. We propose a three-phase sub-syllable segmentation approach. In contrast with Hirose[4] and Sun[3], our approach bases segmentation on both amplitude and pitch information related to the articulatory effort of the speaker during tone production. Furthermore, we identify this segmentation based on the convex hull of the amplitude-frequency contour. After we transform the tone contour into a segmented sequence, we run a Gaussian SVM to perform our tone recognition experiment. Our results show that this approach achieves a 15% accuracy improvement over the tone recognition approach without amplitude segmentation and features. The overall accuracy is comparable to the context-independent result in [5].

2. Tone segmentation and feature representation

2.1. Segmentation

Due to coarticulation effects, we believe that only certain portions of the surface tone realization satisfy the definitions of the underlying tone targets, analogous to tone nuclei in [4]. The other segments only contain coarticulatory transitions with previous or following tone objects. In order to separate the regions where the tone object is well-approximated from adjacent coarticulatory intervals, we propose the following three-phase segmentation according to the role of each segment in tone production:

1. **Onset** This region captures the carryover coarticulation from previous tone to the tone of the current syllable.
2. **Middle** The pitch contour in this region is expected to closely approximate the ideal tone shape.
3. **Offset** This region contains the portion of the contour corresponding to anticipatory coarticulation.

2.2. Segmentation from convex hull plot

In this section, we will generate an approximation of the above three-phase segmentation from the convex hull plot of the amplitude-pitch contour. The changes in direction of the amplitude-pitch convex hull plot correspond to variations in pitch and corresponding amplitude. Given the amplitude-pitch plot, we compute its convex hull. We obtain the three-phase segmentation by locating and merging those convex hull edges into no more than three groups which are adjacent and similar enough, specifically where the ratio of the slopes is within a factor of 10 or the slopes have the same signs. Each interval denoted by one convex hull edge can only belong to one phase.

Figure 1 shows an example of the segmentation of a second tone. The pitch contour appears in the upper left, and the amplitude contour in the upper right. The amplitude-frequency plot and its convex hull appear in the bottom left, with the resulting segmentation in the lower right. The segmentation yields three sub-syllable regions. The onset phase, marked with squares, and corresponds to low amplitude and slightly rising frequency. The middle phase, marked with circles, corresponds to the portion of the tone with both elevated amplitude and rising frequency. The offset phase, marked with diamonds, corresponds to the decrease in amplitude level and frequency. In this example, the region of high amplitude corresponds to the rising pitch of the second tone's canonical contour. Moreover, it captures more information than pitch alone by showing the drop in amplitude 30ms prior to perceptible pitch lowering.

We also performed a side experiment to explore the comparable convex hull-based segmentation directly on the pitch contour itself. We found that this approach often undersegmented the tone, missing either onset or offset regions. Figure 2 is the pitch-only segmentation for the same tone sample shown with the amplitude-frequency based segmentation in Figure 1. Here, only two points are identified as the offset region, but the approach is not sensitive enough to differentiate further. In contrast, the combination of amplitude and frequency highlights these transitions.

3. Experiment

3.1. Data preparation

We extract our dataset from Voice of American Mandarin broadcast news.[1] The audio material includes news reports by various speakers recorded in May of 1998. These recordings were automatically force-aligned to so-called "anchor scripts", using the language porting functionality of the CUsonic speech recognizer [10]. This alignment employed a large pinyin pronunciation lexicon and a manually constructed mapping from pinyin to ARPABET. In addition, this alignment provided word, syllable, and phoneme position used in the subsequent experiments. Severe errors due to mistranscription were manually corrected, as were some tone errors due to speaker variation. Finally, tone sandhi rules were applied. We perform a four-way classification on our dataset, corresponding to four canonical Mandarin Chinese tones. [2]

3.2. Feature extraction and representation

In feature extraction, we first obtain the pitch contour and amplitude contour using Praat [11] "To pitch... " and "To intensity..." commands. Under the syllable segmentation obtained above, we further divide the contour according to our sub-syllable segmentation. Then we represent each segment with eight features of four different dimensions: frequency, amplitude, segment position and energy. For frequency and amplitude, we compute both mean and standard deviation. Similarly, we consider not only the length of a given segment but also its location within a syllable. As for energy, we consider not only the current segment energy but also the total energy from the beginning of the syllable to the end of the current segment. [3]

3.3. Experiment Logistics

For tones with missing phases, for example when the onset or offset phase is absent, we linearly interpolate our samples by adding several points for each missing segment. Hence, for each syllable we construct a three-phase feature vector where each phase contributes eight features for a full vector of twenty-four features. Given this syllable representation, we train an SVM with Gaussian kernel ([12]) $(K(x, y) = exp(-\frac{\|x-y\|^2}{2\sigma^2})$ with fixed $\sigma = 1$) on 10000 samples and then test it on a distinct test set of 500 samples.

3.4. Experiment contrast

We perform contrastive experiments along the following three different dimensions. The first comparison is across segmentation conditions. Two experiments employ feature vectors from multiple segments (interpolated) and two experiments consider only a single segment. In the single segment experiments, tone recognition is tested based on two different strategies for extracting the region with most canonical tone region, or tone nucleus. One approach is to select the segment with maximum energy from the multiple segmentation, which is expected to correspond with the middle phase segment. The other approach assumes, following [1] and [3] that the tone target is best approximated in the second half of the syllable and extracts features only for this region.

In the second comparison, we further consider different feature representations. These experiments contrast classification accuracy using both amplitude and frequency features and using frequency and duration features, excluding amplitude. Motivated by the 5-level tone templates in Stem ML, we perform a third comparison between unbinned data and binned data which maps all values to 10 uniform width bins between 0 and 1.

4. Results and Discussion

4.1. Results

Table 2 presents the results of tone recognition using only a single sub-segment, identified either purely by position in syllable (Second Half) or by energy rank within our amplitude-frequency

[1] Available from http://ldc.upenn.edu

[2] We exclude neutral tone which appears only on unstressed syllables and lacks a canonical pitch contour.

[3] We multiply all pitch and amplitude features by -1 if the slopes of pitch and amplitude contour are not the same on the current segment. E.g the pitch increases with amplitude decreases and vice versa

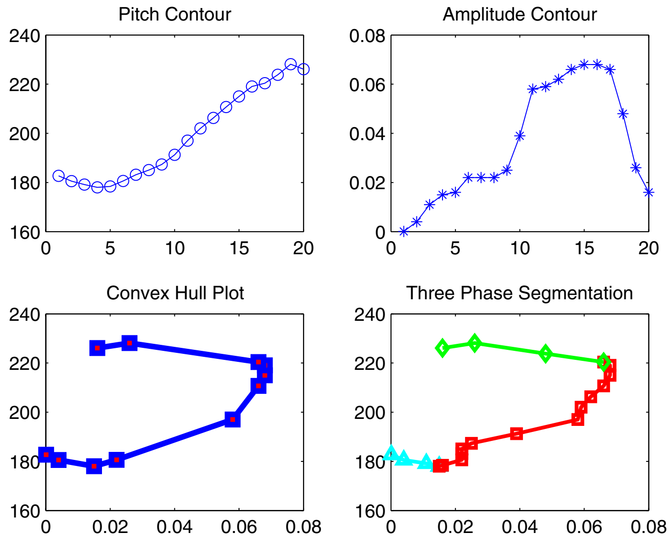

Figure 1: Three-phase segmentation of Second Tone. Pitch contour (top left); Amplitude contour(top right); Convex hull of amplitude-frequency plot (bottom left); Final segmentation (bottom right)

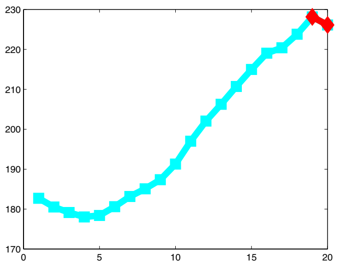

Figure 2: Pitch only convex hull segmentation of a second tone

based segmentation. By exploiting the sub-syllable segmentation and relative energy information to identify the region of maximum interest, this approach substantially outperforms the simple position based strategy.

Table 1 presents the results for tone recognition using the full three-phase feature vector, varying the feature types, with and without amplitude, and the feature representations, binned and unbinned. The best results are obtained using both amplitude and pitch features, reaching 68.2% accuracy, and binned features yielded slightly better results than unbinned.

The above results indicate the utility of the amplitude-frequency based segmentation for tone recognition. While the best results use the full vector from three-phase segmentation, it is interesting to observe that the best single segment results using our hypothesized tone nucleus extraction achieve only slightly poorer performance. This level of accuracy suggests that the nucleus is in fact capturing much of the information required for effective tone recognition. While pitch naturally provides much of the information for tone recognition, amplitude both in the form of features for classification and for tone nucleus identification and segmentation plays an important role.

	Amplitude	No Amplitude Features
Binned	68.2%	66.2%
Unbinned	66.4%	65.8%

Table 1: Tone recognition using interpolated three-phase segmentation with and without amplitude features

	Second Half	Max Energy (Interpolated Segmentation)
Binned	44.2%	62.6%
Unbinned	41.1%	65.4%

Table 2: Tone recognition using only a single segment

4.2. Confusion analysis

We provide the confusion matrix for the interpolated three-phase experiment which yielded the highest accuracy, 68.2%, on our test set. As shown in Table 3, most of the confusion occurs between third and fourth tone. There are the possible explanations for this confusion. First, many third tone samples in our corpus are topic initial syllables with higher pitch values than the normal third tone. Consequently, these third tones overlap the pitch range of the normal fourth tones. Second, in continuous speech, the third tone often lacks its distinctive final rise, further enhancing its similarity to the fourth tone. Finally, the falling tone (Tone 4) is the most frequent tone in our dataset, while the third tone is the least frequent, biasing the classifier in favor of Tone 4.

	True 1	True 2	True 3	True 4
Classified 1	98	22	10	16
Classified 2	8	75	7	7
Classified 3	7	14	35	14
Classified 4	14	9	31	133

Table 3: Confusion Matrix

5. Conclusion and future work

Recent research has demonstrated the utility of amplitude in human tone perception, even in the absence of pitch information. In contrast to the approach of [4] using k-means clustering of pitch contour to determine the tone nuclei, our approach exploits both amplitude information and the standard pitch features in a graphical framework to identify the key regions of the syllable for tone recognition. It overcomes the challenges to tone recognition by tonal coarticulation and enhances the feature space representation. Employing a segmentation based on the convex hull of the amplitude frequency plot allows us to distinguish coarticulatory regions from those approaching canonical tone form. Selection of the maximum energy region from that segmentation captures much of the information required for tone recognition, outperforming a simple time-based strategy, and use of the full feature vector across the segmentation further enhances accuracy.

5.1. Future work

In future work, we will extend our tone recognition approach to perform sequential tone recognition across an utterance. Therefore, we will employ sequential discriminative classification techniques such as Hidden Markov Support Vector Machines (HMSVM)[13] to model dependencies between adjacent tones. In these experiments, we will model both the phase and the target tone simultaneously and exploit a tri-gram model for sequential dependencies. This framework will allow enhanced sequence modelling and improved tone recognition.

6. Acknowledgement

This work was supported by NSF Grant #0414919

7. References

[1] Yi Xu, "Transimitting tone and intonation simultaneoulsy:the parallel encoding and target approximation (PENTA) model," *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, pp. 215–220, 2004.

[2] Greg Kochanski and Chilin Shih, "Prosody modelling with soft templates," *Speech Communication*, vol. 39, no. 3-4, pp. 311–352, 2003.

[3] Xuejing Sun, *The determination, analysis, and synthesis of fundamental frequency*, Ph.D. thesis, Northwester University, 2002.

[4] Jinsong Zhang and Keikichi Hirose, "Tone nucleus modelling for chinese lexical tone recognition," *Speech Communication*, vol. 42, pp. 447–466, 2005.

[5] Gina-Anne Levow, "Context in multi-lingual tone and pitch accent recognition," *International Conference on Speech Communication and Technology*, 2005.

[6] Chao Wang and Stephanie Seneff, "Improving tone recognition by normalizaing for coarticulation and intonation effects," *International Conference on Spoken Language Processing*, 2000.

[7] Hansjorg Mixdorff, Yu Hu, and Denis Burnham, "Visual cues in mandarin tone perception," *International Conference on Speech Communication and Technology*, pp. 405–408, 2005.

[8] Yi Xu and D.H Whalen, "Information for mandarin tones in the amplitude contour and in brief segments," *Phonetica*, vol. 49, pp. 25–47, 1992.

[9] Fan-Gang Zeng, Kaibao Nie, Ginger S. Stickney, Ying-Yee Kong, Michael Vongphoe, Ashish Bhargave, Chaogang Wei, and Keli Cao, "Speech recognition with amplitude and frequency modulations," *Proceeding of National Academy of Sciences*, vol. 102, no. 7, 2005.

[10] Rubén San-segundo, Bryan Pellom, and Wayne Ward, "Confidence measures for dialogue management in the cu communicator system," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2000.

[11] Paul Boersma and David Weenink, "Praat: doing phonetics by computer," *Glot International*, vol. 5(9/10), pp. 341–345, 2001.

[12] Vladimir N.Vapnik, *The Nature of Statistical Learning Theory*, Springer Series in Statistics. Springer, 1995.

[13] Yasemin Altun, Ioannis Tsochantaridis, and Thomas Hofmann, "Hidden markov support vector machines," *International Conference on Machine Learning*, 2003.