IMPROVED TONE RECOGNITION BY NORMALIZING FOR COARTICULATION AND INTONATION EFFECTS[1]

Chao Wang and Stephanie Seneff

Spoken Language Systems Group Laboratory for Computer Science
Massachusetts Institute of Technology Cambridge, Massachusetts 02139 USA
{wangc,seneff}@sls.lcs.mit.edu

ABSTRACT

We have previously demonstrated that tone modeling improved speech recognition on a digit corpus [7]. In this work, we further improve tone recognition by normalizing for both *tone coarticulation* and *intonation* effects. The tone classification errors on continuous digit strings were reduced by 26.1% from the baseline, when the effects of F_0 downdrift, phrase boundary and tone coarticulation were normalized. We also applied the same approach to conversational speech from the YINHE domain [6], and obtained similar improvements. The word error rate on *spontaneous* YINHE data was reduced by 16.5% when a simple four-tone model was applied to resort recognizer *10*-best outputs.

1. INTRODUCTION

Tone is a natural target for prosodic modeling in tonal languages, because of its important role in lexical access. There are four lexical tones in Mandarin Chinese, each corresponding to a canonical F_0 contour pattern: "high-level", "high-rising", "low-dipping" and "high-falling". However, tones in continuous speech can vary dramatically from the canonical form, due to coarticulatory effects from surrounding tones, as well as influences from intonation.

The problem of tone coarticulation has been studied by a number of researchers. Shen [4] studied all possible combinations of tones of Mandarin on /pa pa pa/ tri-syllables, and found that not only the onset and offset values but also the overall heights of a tone were affected; and the coarticulatory effects are bi-directional and symmetric. Xu [8] conducted a perceptual study of coarticulated tones and found that human performance on tone identification was highly dependent on the availability of original tonal context when the context was "conflicting" with the tone. Xu [9] also studied F_0 contours of Mandarin bi-syllables /ma ma/ embedded in a number of carrier sentences, and found asymmetrical bi-directional coarticulatory effects in terms of F_0 onset and offset changes.

The interaction between intonation and tone in Mandarin is not yet well understood. A study conducted by Shen [3] on a small set of read utterances found that intonation perturbed both the shape and scale of a given tone. For example, interrogative intonation raises the tone value of the sentence-final syllable as well as the overall pitch level, and tone 1 rises slightly in sentence initial position and falls slightly in sentence final position under statement intonation, etc. However, it was concluded that the basic tone shape is preserved, i.e., tone 4 did not become falling-rising under question intonation, as suggested by Chao [1].

We have implemented a basic tone classification system and demonstrated improved speech recognition by adding tone models for a digit corpus [7], which contains random digit strings of 5 to 10 digits and 9-digit phone numbers. In this work, we will study and characterize tone coarticulation effects as well as the interaction between tone and intonation, using statistical methods and a large corpus of speech data. In addition to the digit data, we will also apply our approach to more linguistically rich data from the YINHE domain [6], which contains read queries as well as *spontaneous* utterances of users interacting with a conversational system for flight, weather, and local city-guide information. We will improve tone classification performance by normalizing for these influences, and apply tone models to improve recognition performance of spontaneous speech. In the following, we describe our modelling approach, and provide some experimental results on tone classification and speech recognition.

2. F_0 DOWNDRIFT

According to the phonological approach to intonation [2], the intonation contour is a string of pitch accents and boundary tones, and there is an overall downstep trend of the F_0 level of the pitch accents. We will only account for the downdrift and boundary influences in this study. In this section, we focus on modeling the F_0 downdrift. The boundary effects will be characterized as a context to tone and discussed in the next section.

It is generally agreed that tonal languages may make use of a limited amount of superimposed intonation. However, because both tone and intonation are manifested as F_0 movements, it is difficult to separate the two aspects in the physical signal. Wang [5] adapted Fujisaki's model for the F_0 contour of Mandarin Chinese, in which the F_0 contour (in logarithmic form) was represented as the sum of a phrase component and a tone component, each being approximated by the response of a second-order linear system to the respective phrase or tone commands. The model was applied in a tone recognition task for Chinese four-syllable idioms, for which a single phrase is assumed.

Given digit data, we assume that all utterances have a similar underlying intonation contour. Thus a pitch contour can be viewed as a "constant" intonation component with additive "random" perturbations caused by tones. We can use an averaging ap-

[1]This work was supported by the National Science Foundation under Grant No. IRI-9618731; and by DARPA under contract N66001-99-1-8904, monitored through Naval Command, Control, and Ocean Surveillance Center.

proach to smooth out the "random" variations due to tones and obtain the average as the underlying intonation contour.

We tested our hypothesis by plotting the F_0 contours of all digit data, grouped by random digit strings and phone numbers, in Figure 1. The time scale of each utterance is normalized by the utterance duration, so that utterances of different length can be aligned in time. It is obvious from the plot that there is a steady downdrift of the mean pitch contour, although the slope for the downdrift trend is slightly different for random digit strings and phone numbers. The F_0 contour plot of phone numbers also reveals a more detailed phrase structure corresponding to the habitual way of grouping digits in phone numbers ("xx-xxx-xxxx"), which is most obviously marked by a sharp drop of F_0 at phrase boundaries. We believe that a random digit string also has similar behavior in its F_0 contour. The absence of such evidence from the plot is due to the "randomized" positions of the phrase boundaries in the time-normalized F_0 contour; thus the "averaging" also smoothed out the phrase boundaries.

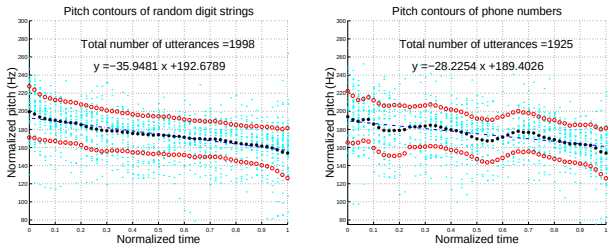

Figure 1: Pitch contours of random digit strings and phone numbers. The starred line represents the mean pitch contour, with the upper and lower circled lines for standard deviation. The dashed line is the linear regression line for the average F_0 contour.

To determine if the downdrift slope is affected by the utterance duration or phrase structure, we grouped the random digit strings according to their number of syllables and phrases, and obtained average F_0 slopes for each subset, as plotted in Figure 2. We were not able to obtain the slopes for utterances with 3 or more phrases, because of sparse data problem. From the plot we can see that the slopes for 2-phrase utterances are consistently smaller than their 1-phrase counterparts, suggesting that the F_0 base is raised after a pause. The trend regarding the number of syllables per utterance is not very obvious.

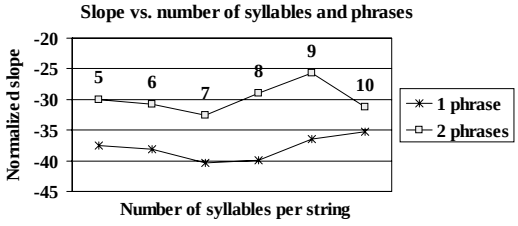

Figure 2: Downdrift slope of random digit strings grouped by number of syllables and phrases.

It is unclear if the downdrift factor can be modeled as a constant for the various types of utterances in the YINHE domain. We examined that by comparing the mean F_0 contour for different utterance types. The utterances were labelled manually using four categories, including declarative, command, wh-question, and particle-question, which is similar to the yes-no question in English. As indicated in Figure 3, there are some differences in the F_0 slope, with wh-questions having the sharpest drop and commands having the least. We believe that the small slope in the "command" utterances is an artifact caused by the biased tone content, i.e., a large portion of the data correspond to "fan3 hui2" (go back), causing the F_0 contour to rise at the end; however, the relatively short duration of this type of utterance might also play a role. A scatter plot of the slope vs. duration for each utterance reveals a slight correlation between the two. However, considering that the across-type differences are relatively small (especially when compared with the large across-utterance differences observed in the data), we are inclined to use a single downdrift factor for all the data. We realize, however, that this is somewhat inadequate for the YINHE data, as indicated by our experimental results.

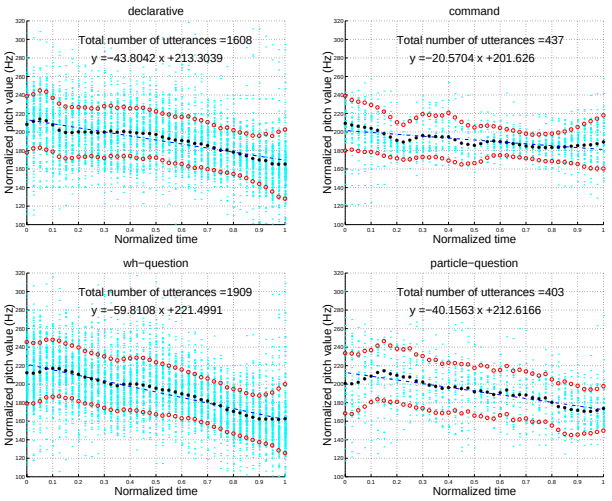

Figure 3: Pitch contours of YINHE utterances grouped by utterance type. The starred line represents the mean pitch contour, with the upper and lower circled lines for standard deviation. The linear regression coefficients are also shown in the figures.

3. PHRASE BOUNDARY AND COARTICULATION EFFECTS

We can use the statistical difference of context-independent (ci-) models and context-dependent (cd-) models to characterize the tonal coarticulatory effects as well as the influence of phrase boundary to tones. Our system uses Legendre transformation coefficients of the F_0 contour as tone features, the first two of which correspond to the level and slope of the tone contour. We will use these two measurements as examples to demonstrate the contextual influences. The means of these two measurements are shown in Figure 4 for all right- and left- cd-models. The ci-model statistics are also displayed for comparison.

3.1. Phrase Boundary

We distinguish between internal phrase boundary (denoted by "<>") and the utterance end (denoted by "end"). There exist

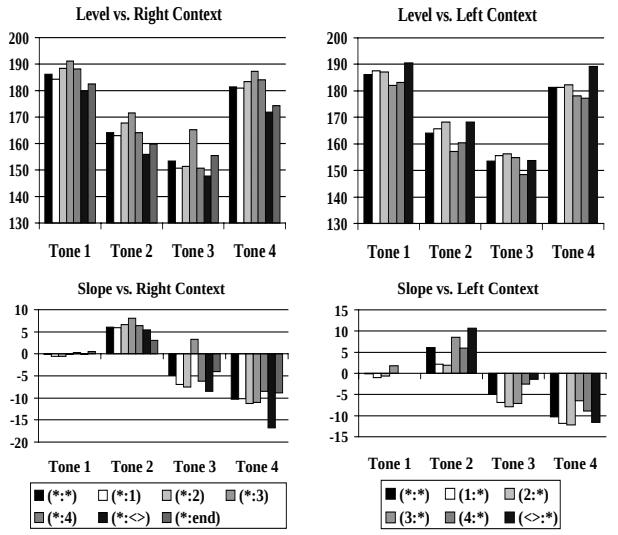

Figure 4: Mean F_0 level and slope of four tone models conditioned on the right context (left side, columns 2 to 7) and left context (right side, columns 2 to 6). Mean of context-independent models shown in column 1.

relatively large discrepancies between tonal statistics before the internal boundary and the utterance final boundary. We suspect that this might be an artifact caused by poor pitch tracking performance at the utterance end due to glottalization, which frequently resulted in a flat pitch contour with doubled pitch value. In the following, we will focus on discussing the effects of the internal phrase boundary.

As shown by the left-side plots in Figure 4, the tones at phrase-final syllable position (indicated by a right context of "<>") have a relatively lower F_0 level, a larger falling slope, and a smaller rising slope. This seems to suggest that the intonation contour falls at the phrase boundary. However, notice that the dropping for tone 4 at a phrase boundary is much larger than that of tone 3, and the decrease of the rising slope of tone 2 is relatively small. This implies that a simple subtraction of the observed F_0 dropping at phrase boundary from the F_0 contour will not do very well in removing these influences.

As shown by the right-side plots in Figure 4, the tones at phrase-initial syllable position (indicated by a left context of "<>") generally have a higher F_0, a larger rising slope, and a smaller falling slope, except for tone 4. Again, the magnitude of these relative differences also varies.

Overall, the data seem to support the argument that intonation modifies the shape of tones; however, the basic pattern remains intact, as concluded by Shen [3].

3.2. Tone Coarticulation Effects

The left two plots in Figure 4 demonstrate the effects of the right tonal context on the distribution of the F_0 level and slope for the four lexical tones, shown in columns two to five. The ci-model statistics are shown in column one to facilitate comparison. The following observations can be made from these plots:

- Tone 3 preceding tone 3 has average F_0 level and slope similar to those of tone 2. This is due to the well-known tone-sandhi rule that tone 3 before tone 3 becomes tone 2.
- All other tones before tone 3 also have a much higher level than when they precede the other tones, and all tones are slightly lowered before tone 1.

The right two plots in Figure 4 demonstrate the effects of the left tonal context on the distribution of the F_0 level and slope for the four lexical tones, shown in columns two to five. The ci-model statistics are shown in column one. The following observations can easily be made from the plots:

- All tones after tone 1 and tone 2 (high offset) have a higher level than after tone 3 and tone 4 (low offset). The seeming exception of tone 3 after tone 3 can be explained by the "$33 \rightarrow 23$" tone-sandhi rule.
- Tone 2 after tones 1 and 2 has a much smaller rising slope than after tones 3 and 4; while tone 4 after tones 1 and 2 has a much larger falling slope than after tones 3 and 4.
- Tone 3 after tone 1, tone 2, and tone 3 (effectively tone 2) has a much larger falling slope than after tone 4.

These observations seem to suggest that anticipatory effects are dissimilatory, and the carry-over effects are assimilatory in nature. For example, tones preceding low tone (tone 3) raise their F_0 level to contrast with the low F_0 and tones preceding high tone (tone 1) lower their F_0 level to contrast with the high F_0, however, tones after high offset tones (tone 1 and tone 2) tend to have higher F_0 level, smaller rising slope but larger falling slope, and tones after low offset tones (tone 3 and tone 4) have exactly the opposite behavior. It also seems that the carry-over effects are larger than anticipatory effects, as indicated by larger differences among the left- cd-models for the same tone. These observations support Xu's conclusions in [9].

4. EXPERIMENTAL RESULTS

The baseline tone recognition system is described in detail in [7]. The test data consist of 355 digit strings, as well as 194 spontaneous and 206 read YINHE utterances from 6 speakers. The tone classification results are obtained for the three sets respectively for comparison of different speaking styles. Neutral tones are excluded, because they contribute to a large percentage of classification errors, and perform worse in resorting the recognizer output. In the following, we describe the method to incorporate the models in our tone system, followed by a summary of results.

4.1. F_0 Downdrift Normalization

We started by modeling the F_0 downdrift as a straight line for the digit and the YINHE data. The parameters are estimated by linear regression analysis of the mean F_0 contour. We subtracted this downdrift from each F_0 contour and re-trained tone models. This significantly reduced the tone classification errors, as shown in Tables 1 and 2. A closer examination of the model parameters revealed that the model variances were also greatly reduced.

We also tried various ways to achieve further performance improvement. One attempt is to use more refined phrase models instead of a sentence level model, motivated by Figure 2. We tested this approach on phone numbers, which usually contain three phrases in each utterance. A regression line for each phrase is obtained from the mean F_0 contour of all training data. However, this refinement did not yield significant improvement over the simple sentence model. We also tried various regression analyses for each *individual* utterance's F_0 contour to approximate the intonation component. However, the tone classification performance degraded. We observed that the resulting intonation curve using this approach follows the F_0 contour too closely, thus consuming part of the contribution from tones. This approach is also not robust to errors in pitch extraction and segmental alignment, which are frequent in telephone speech.

4.2. Context Normalization

We used a "corrective" approach to account for different tonal contexts, replacing the context-dependent tone models used in the baseline system. Specifically, we alter the F_0 contour of each tone according to its contexts to compensate for the coarticulatory effects, manifested as the differences between the ci-models and cd-models. New ci-models are then trained from those corrected F_0 contours. We performed correction for the F_0 average and slope. We found that the variances of the new models were significantly reduced on those two dimensions, and the classification errors are further reduced.

4.3. Summary of Results

Tables 1 and 2 summarize the tone classification results for the three data sets. The baseline performance has an 18.4% error rate for digit data, 32.5% for read YINHE data, and 35.2% for spontaneous data, consistent with the complexity of each data set. The normalization schemes demonstrate performance improvements for all three sets, as shown in the tables.

System Configuration	Classification Error Rate(%)	Relative Reduction(%)
Baseline	18.4	-
+ Intonation	15.9	13.5
+ Context	14.7	20.1
+ Both	13.6	26.1

Table 1: Four tone classification results on the digit data.

System Configuration	Read		Spontaneous	
	ER(%)	Rel.(%)	ER(%)	Rel.(%)
Baseline	32.4	-	35.2	-
+ Intonation	29.3	9.6	33.1	6.0
+ Both	27.6	14.8	31.2	11.4

Table 2: Four tone classification results on spontaneous and read YINHE data (neutral tone excluded).

We applied the tone models to resort the *10*-best outputs of the YINHE recognizer. The read YINHE data were used to optimize the relative weight of the tone score contribution, and the speech recognition performance is reported on the spontaneous data. Application of various tone models all reduced the word and sentence error rates, as shown in Table 3. However, the improved classification performance with intonation and context normalization did not translate into consistent recognition improvements. Further study of the data needs to be done to find an explanation.

System	Sub.	Ins.	Del.	WER	SER
No Tone	6.2	1.4	0.8	8.5	29.4
Baseline	5.1	1.4	0.6	7.1	27.3
+ Intonation	5.3	1.5	0.6	7.5	28.4
+ Both	5.0	1.3	0.6	6.9	26.8

Table 3: Recognition results (in percentage) on spontaneous YINHE data with no tone models and various tone models.

5. SUMMARY AND DISCUSSION

In addition to tone coarticulation effects, we have accounted for two factors in the intonation component, i.e., the downdrift of F_0 throughout an utterance and the phrase boundaries. We believe that using the mean F_0 contour over a pool of similar utterances to estimate the downdrift is more robust than using the individual contours. We also showed that the phrase boundary, specifically, a falling boundary, does not simply superimpose a large drop of F_0 on all tones, as indicated by our context-dependent tone models. We have not studied the influence of pitch accents in our study. This could potentially be done by conditioning on the "pitch accent" property of the underlying syllable in the context-dependent models. However, this requires manually labelled data, before an automatic method can be trained. Currently our context normalization is implemented as a pre-processing of the F_0 contour given a tone transcription. We are in the process of refining our method to remove this dependency by obtaining the context from the recognition *N*-best outputs instead.

6. REFERENCES

1. Y-R. Chao, "Tone and intonation in Chinese," in *145th meeting of the American Oriental Society*, pp. 121-134, 1933.
2. R. D. Ladd, *Intonational Phonology*. Cambridge University Press, 1996.
3. X-N. Shen, "Interplay of the four citation tones and intonation in Mandarin Chinese," in *Journal of Chinese Linguistics*, vol. 17, no. 1, pp. 61-74, 1989.
4. X-N. Shen, "Tonal coarticulation in Mandarin," in *Journal of Phonetics*, vol. 18, pp. 281-295, 1990.
5. C.-F. Wang, H. Fujisaki, and K. Hirose, "Chinese four tone recognition based on the model for process of generating F_0 contours of sentences," in *Proc. ICSLP'90*, Kobe, Japan, pp. 221-224, 1990.
6. C. Wang, J. Glass, H. Meng, J. Polifroni, S. Seneff and V. Zue, "Yinhe: A Mandarin Chinese version of the Galaxy system," in *Eurospeech'97*, Rhodes, Greece, pp. 351-354, 1997.
7. C. Wang and S. Seneff, "A study of tones and tempo in continuous Mandarin digit strings and their application in telephone quality speech recognition," in *Proc. ICSLP'98*, Sydney, Australia, pp. 635-638, 1998.
8. Y. Xu, "Production and perception of coarticulated tones," in *Journal of Acoustic Society of America*, vol. 95, no. 4, pp. 2240-2253, 1994.
9. Y. Xu, "Contextual tonal variations in Mandarin," in *Journal of Phonetics*, vol. 25, pp. 62-83, 1997.