Turn-taking in Mandarin Dialogue: Interactions of Tone and Intonation

Gina-Anne Levow
Computer Science Department
University of Chicago
Chicago, IL 60637 USA
levow@cs.uchicago.edu

Abstract

Fluent dialogue requires that speakers successfully negotiate and signal turn-taking. While many cues to turn change have been proposed, especially in multi-modal frameworks, here we focus on the use of prosodic cues to these functions. In particular, we consider the use of prosodic cues in a tone language, Mandarin Chinese, where variations in pitch height and slope additionally serve to determine word meaning. Within a corpus of spontaneous Chinese dialogues, we find that turn-unit final syllables are significantly lower in average pitch and intensity than turn-unit initial syllables in both smooth turn changes and segments ended by speaker overlap. Interruptions are characterized by significant prosodic differences from smooth turn initiations. Furthermore, we demonstrate that these contrasts correspond to an overall lowering across all tones in final position, which largely preserves the relative heights of the lexical tones. In classification tasks, we contrast the use of text and prosodic features. Finally, we demonstrate that, on balanced training and test sets, we can distinguish turn-unit final words from other words at $\approx 93\%$ accuracy and interruptions from smooth turn unit initiations at 62% accuracy.

1 Introduction

Fluent dialogues require effective turn transitions between speakers. Research in turn-taking, typified by (Duncan, 1974), posits several key signals for turn-taking, including a turn-change signal which offers to cede the floor, speaker-state signal which indicates taking the floor, within-turn signal, and continuation signals. This process fundamentally requires cooperation between participants both to produce the contextually appropriate signals and to interpret those of their interlocutor. These analyses have proposed a wide range of cues to turn status, ranging from gaze and gesture in a multi-modal context to prosodic cues including pitch, intensity, and duration as well as lexical and syntactic cues.

Much of this fundamental research as well as computational implementations have focused on English, a language with well-studied intonational sentence and discourse structure. A substantial body of work has identified sentence-like units as well as fragments and repairs in conversational speech, including (Ostendorf, forthcoming; Liu et al., 2004; Shriberg et al., 2001) These approaches have employed lexical and prosodic cues in diverse frameworks, including Hidden Markov Models employing decision trees and hidden state language models, neural networks, and maximum entropy models. (Shriberg et al., 2001) identified jump-in points and jump-in words in multi-party meeting speech using prosodic and language model features at accuracies of 65 and 77% under equal priors. Furthermore, (Ward and Tsukahara, 2000) demonstrated that backchannels occurred at predictable points

with specific prosodic characteristics in both English and Japanese.

In the current paper, we consider the interaction of potential prosodic intonational cues related to turn-taking with the realization of lexical tone in a tone language, Putonghua or Mandarin Chinese. Mandarin employs four canonical lexical tones distinguished by pitch height and pitch contour: high level, mid-rising, low falling-rising, and high falling. We explore whether prosodic features are also employed in turn-taking behavior in this language and whether the forms are comparable to those employed in languages with lexical tone. We demonstrate that intonational cues quite similar to those in English are also employed in Chinese with lower pitch and intensity at ends of turn units than at the beginnings of those turn units. Interruptions likewise are distinguished from smooth turn transitions by prosodic means, including greater pitch elevation. We demonstrate how these changes interact with lexical tone by substantial lowering of average pitch height across all tones in final positions and contrast pitch contours in final and non-final positions. Finally, these cues in conjunction with silence and durational features can be employed to distinguish turn-unit final words from non-final words in the dialogue and words that initiate interruptions from those which start smoother turn transitions.

In the remainder of the paper we will briefly describe the data set employed in these experiments and the basic extraction of prosodic features (Section 2). We then present acoustic analyses contrasting turn unit final and turn unit initial syllables under different turn transition types (Section 3). We will describe the impact of these intonational cues on the realization of lexical tone (Section 4). Finally we will apply these prosodic contrasts to enable classification of words for finality and interruption status (Section 5).

2 Experimental Data

The data in this study was drawn from the Taiwanese Putonghua Speech Corpus[1]. The materials chosen include 5 spontaneous dialogues by Taiwanese speakers of Mandarin, seven females and three males. The dialogues, averaging 20 minutes in duration, were recorded on two channels, one per speaker, in a quiet room and digitized at 16KHz sampling. The recordings were later manually transcribed and segmented into words; turn-beginnings and overlaps were time-stamped. The manual word segmentation was based on both syntax and phonology according to a methodology described in detail in (Duanmu, 1996).

2.1 Prosodic Features

For the subsequent analysis, the conversations were divided into chunks based on the turn and overlap time-stamps. Using a Chinese character-to-pinyin dictionary and a hand-constructed mapping of pinyin sequences to ARPABET phonemes, the transcribed text was then force-aligned to the corresponding audio segments using the language porting mechanism in the University of Colorado Sonic speech recognizer (Pellom et al., 2001). The resulting alignment provided phone, syllable, and word durations as well as silence positions and durations.

Pitch and intensity values for voiced regions were computed using the functions ”To Pitch” and ”To Intensity” in the freely available Praat acoustic analysis software package(Boersma, 2001). We then computed normalized pitch and intensity values based on log-scaled z-score normalization of each conversation side. Based on the above alignment, we then computed maximum and mean pitch and intensity values for each syllable and word for all voiced regions. Given the presence of lexical tone, we extracted five points evenly distributed across the ”final” region of the syllable, excluding the initial consonant, if any. We then estimated the linear syllable slope based on the latter half of this region in which the effects of tonal coarticulation are likely to be minimized under the pitch target approximation model(Xu, 1997).

3 Acoustic Analysis of Turn-taking

Each of the turn units extracted above was tagged based on its starting and ending conditions as one of four types: smooth, rough, intersmooth, and interrough. ”Smooth” indicates a segment-ending transition from one speaker to another, not caused

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

Position	Start −Overlap or −Spkr Change	Start +Overlap +Spkr Change
End −Overlap	Smooth 1413	Intersmooth 289
End +Overlap	Rough 407	Interrough 68

Table 1: Types of turn units

by the start of overlap with another speaker. By contrast, a rough transition indicates the end of a chunk at the start of overlap with another speaker. The prefix "inter" indicates the turn began with an interruption, identified by overlap with the previous speaker and change of speaker holding the floor. In this class, the new speaker continues to hold the floor after the period of overlap.

We contrast turn unit initial and turn unit final syllables for each type of transition and across all turns. We compare mean pitch and mean intensity in each case. We find in all cases highly significant differences between mean pitch of turn unit initial syllables and mean pitch of final syllables ($p < 0.0001$) as illustrated in Figure 1. Syllables in initial position have much higher log-scaled mean pitch in all conditions. For intensity, we find highly significant differences across all conditions ($p < 0.005$), with initial syllables having higher amplitude than final syllables. These contrasts appear in Figure 2. Furthermore, comparing final intensity of transitions not marked by the start of overlap with the intensity of the final pre-overlap syllable in a transition caused by overlap, we find significantly higher normalized mean intensity in all rough transitions relative to others. In contrast, comparable differences in pitch do not reach significance.

Finally we compare smooth turn unit initiations ("smooth") to successful interruptions ("interrough","intersmooth"), contrasting initial syllables in each class. Here we find that both normalized mean pitch (Figure 3) and normalized mean intensity (Figure 4) in turn unit initial syllables are significantly higher in interruptions than in "smooth" turn transitions.[2] Speakers use these

[2]If one compares both "smooth" and "rough" transitions to "intersmooth" and "interrough" transitions, initial syllables are significantly higher in pitch for the interruption

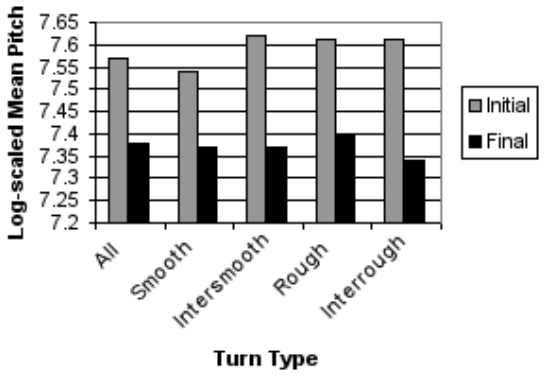

Figure 1: Pitch contrasts between syllables in initial and final position across turn types. Values for initial position are in grey, final position in black.

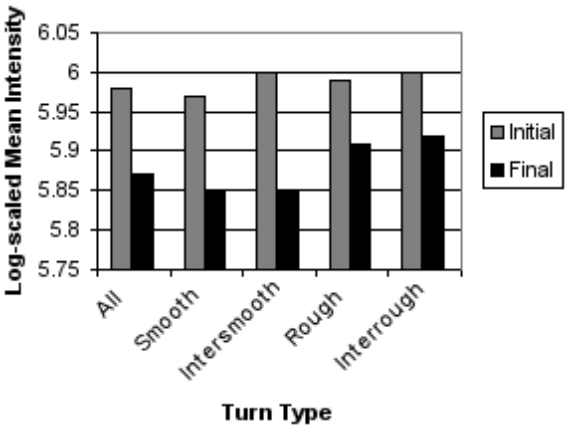

Figure 2: Intensity contrasts between syllables in initial and final position across turn types. Values for initial position are in grey, final position in black.

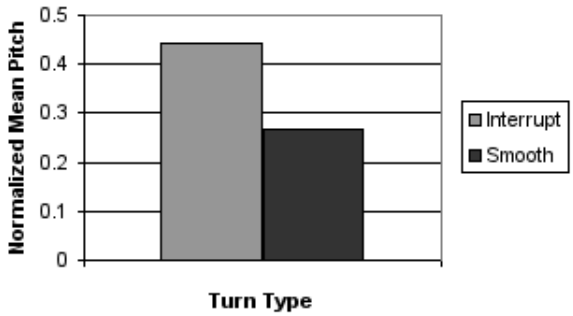

Figure 3: Pitch contrasts between initial syllables in smooth turn transitions and interruptions. Values for smooth transitions are in black, interruptions in grey.

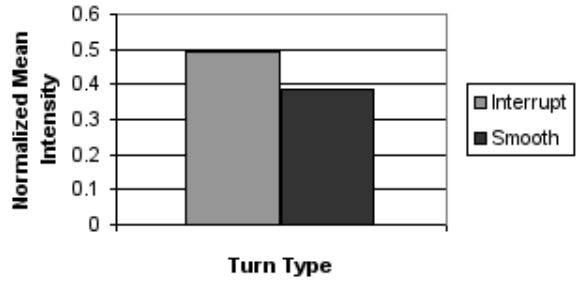

Figure 4: Intensity contrasts between initial syllables in smooth turn transitions and interruptions. Values for smooth transitions are in black, interruptions in grey.

prosodic cues to take the floor by interruption.

These descriptive analyses demonstrate that intonational cues to turn-taking do play a role in a tone language. Not only does intensity play a significant role, but pitch also is employed to distinguish initiation and finality, in spite of its concurrent use in determining lexical identity. In the following section, we describe the effects on tone height and tone shape caused by these broader intonational phenomena.

4 Tone and Intonation

We have determined that syllables in turn unit final position have dramatically reduced average pitch relative to those in turn unit initial position, and these contrasts can serve to signal turn-change and speaker change as suggested by (Duncan, 1974). How do these changes interact with lexical identity and lexical tone? Since tone operates on a syllable in Chinese, we consider the average pitch and tone contours of syllables in final and non-final position. We find that average pitch for all tones is reduced, and relative tone height is largely preserved.[3] Thus a final high tone is readily distinguishable from a final low tone, if the listener can interpret the syllable as turn-final. The contrasts appear in Figure 5.

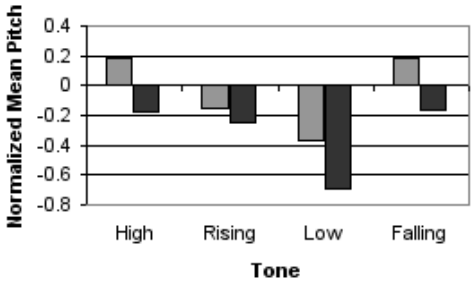

Figure 5: Contrasts in average pitch for the four canonical tones in turn non-final and final positions. Values for non-final positions are in grey, final positions in black.

Turning to tone contour, we find likewise little change between non-final and final contours, with the contours running parallel, but at a much lower pitch.[4] For illustration, mid-rising and high falling tones are shown in Figure 6. Comparable behavior has been observed at other discourse boundaries such as story boundaries in newswire speech. (Levow, 2004).

5 Recognizing Turn Unit Boundaries and Interruptions

Based on the salient contrasts in pitch and intensity observed above, we employ prosodic features both to identify turn boundaries and to distinguish between the start of interruptions and smooth transitions. We further contrast the use of prosodic features with text n-gram features.

classes, but differences for intensity do not reach significance ($p = 0.053$)

[3] This analysis excludes exclamatory and interjective particles.

[4] It is also true that contours do not always match their canonical forms even in non-final position. This variation may be attributed to a combination of tonal coarticulatory effects and the presence of other turn-internal boundaries.

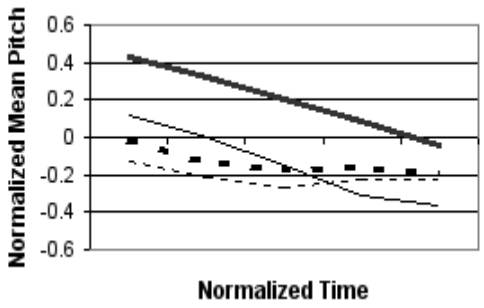

Figure 6: Contrasts in pitch contour for mid-rising and high falling tones in turn non-final and final positions. Values for non-final positions are in heavy lines, final positions in thin lines. Mid-rising tone is in black, dashed lines, high falling in solid lines.

5.1 Classifier Features: Prosodic

The features used in the classifier trained to recognize turn boundaries and turn types fall into two classes: local and contextual. The local features describe the words or syllables themselves, while the contextual features capture contrasts between adjacent words or syllables. The first set of features thus includes the mean pitch and mean intensity for the current word and syllable, the word duration, and the maximum pitch and intensity for the syllable. The second set of features include the length of any following silence and the differences in mean pitch and mean intensity between the current word or syllable and the following word or syllable.

5.2 Classifier Features: Text

For contrastive purposes, we also consider the use of textual features for turn boundary and boundary type classification.[5] Here we exploit syllable and word features, as well as syllable n-gram features. We use the toneless pinyin representation of the current word and the final syllable in each word. Such features aim to capture, for example, question particles that signal the end of a turn. In addition, we extracted the five preceding and five following syllables in the sequence around the current syllable. We then experimented with different window widths for n-gram construction, ranging from one to five, as supported by the classifier described below.

[5] All text features are drawn from the ground truth manual text transcripts.

5.3 Classifiers

We performed experiments with several classifiers: Boostexter (Schapire and Singer, 2000), a well-known implementation of boosted weak learners, multi-class Support Vector Machines with linear kernels (C-C.Cheng and Lin, 2001), an implementation of a margin-based discriminative classifier, and decision trees, implemented by C4.5(Quinlan, 1992). All classifiers yielded comparable results on this classification task. Here we present the results using Boostexter to exploit its support for text features and automatic n-gram feature selection as well as its relative interpretability. We used downsampled balanced training and test sets to enable assessment of the utility of these features for classification and employed 10-fold cross-validation, presenting the average over the runs.

5.3.1 Recognizing Turn Unit Ends

Using the features above we created a set of 1610 turn unit final words and 1610 non-final words. Based on 10-fold cross-validation, using combined text and prosodic features, we obtain an accuracy of 93.1% on this task. The key prosodic features in this classification are silence duration, which is the first feature selected, and maximum intensity. The highest lexical features are preceding 'ta', preceding 'ao', and following 'dui'. If silence features are excluded, classification accuracy drops substantially to 69%, still better than the 50% chance baseline for this set. In this case, syllable mean intensity features become the first selected for classification.

We also consider the relative effectiveness of classifiers based on text, with silence, or prosodic features alone. We find that, when silence duration features are available, both text- and prosodic-based classifiers perform comparably at 93.5% and 93.7% accuracy respectively, near the effectiveness of the combined text and prosodic classifier. However, when silence features are excluded, a greater difference is found between classification based on text features and classification based on prosodic features. Specifically, without silence information, classification based

on text features alone reaches only 59.5%. However, classification based on prosodic features remains somewhat more robust, though still with a substantial drop in accuracy, at 66.5% for prosody only. This finding suggests that although the presence of a longer silence interval is the best cue to finality, additional prosodic features, such as differences in pitch and intensity, concurrently signal the opportunity for another speaker to start a turn. Text features, especially in highly disfluent conversational speech, provide less clear evidence. Results appear in Table 5.3.1.

5.3.2 Recognizing Interruptions

In order to create a comparable context for initial words in interruption and smoothly initiated turns, we reversed the direction of the contextual comparisons, comparing the preceding word features to those of the current word and measuring pre-word silences rather than following silences. Using this configuration, we created a set of 218 interruption initial words and 218 smooth transition initial words, following smooth transitions without overlap. Based on 10-fold cross-validation for this downsampled balanced case, we obtain an accuracy of 62%, relative to a 50% baseline. The best classifiers employed only prosodic features with silence duration and normalized mean word pitch. Addition of text features degrades test set performance as the classifier rapidly overfits to the training materials. If we exclude silence related features, accuracy drops to almost chance.

6 Discussion and Conclusion

We have demonstrated that even in a language that employs lexical tone, cues to turn and speaker status are still carried by prosodic means, including pitch. Specifically, turn unit initial syllables have significantly higher mean pitch and intensity than do final syllables in different turn transition types. The elevation of initial pitch is further enhanced in interruptions, when a new speaker seizes the floor beginning with an overlap of the current speaker. These contrasts are similar to those observed for English, and consistent with signals described in the literature. These changes result in an overall reduction in pitch for syllables in final position across all tones, but relative pitch height employed by lexical tone is preserved in most cases. Finally, we employed these cues to train classifiers to distinguish turn unit final words from other words in the dialogue and to distinguish interruption initial words from initial words in smooth transitions with no overlap.

We further contrasted the utility of text and prosodic features for the identification of turn boundary position and type. We find that silence is the most reliable cue both to identify final turn boundaries and to distinguish types of turn transitions. In conjunction with silence features, text, prosodic, and joint text-prosodic features can fare comparably. However, for turn unit boundaries, the availability of a variety of prosodic features proves to be essential for relatively better classification in the absence of silence information.

In future work, we plan to enhance tone recognition by better contextual modeling that will compensate for the effects of a variety of discourse boundaries, including turn and topic, on tone realization. We also plan to embed turn-based classification in a sequential discriminative model to further facilitate integration of prosodic and lexical features as well as sequence constraints on turn structure.

Acknowledgments

We would like to thank the developers of the Taiwanese Putonghua corpus and the Linguistic Data Consortium for the provision of these resources. This work was supported by NSF Grant 0414919.

References

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9–10):341–345.

C-C.Cheng and C-J. Lin. 2001. LIBSVM:a library for support vector machines. Software available at: http://www.csie.ntu.edu.tw/ cjlin/libsvm.

San Duanmu. 1996. Notes on word boundaries. Taiwanese Putonghua Corpus Documentation.

S. Duncan, 1974. *Some signals and rules for taking speaking turns in conversations*, pages 298–311.

Gina-Anne Levow. 2004. Prosody-based topic segmentation for mandarin broadcast news. In *Proceedings of HLT-NAACL 2004, Companion Volume*, pages 137–140.

	Prosody Only	Text	Prosody + Text
With silence	93.7%	93.5%	93.1%
Without silence	66.5%	59.5%	69%

Table 2: Recognition of turn unit final vs. non-final words

Y. Liu, A. Stolcke, E. Shriberg, and M. Harper. 2004. Comparing and combining generative and posterior probability models: Some advances in sentence boundary detection in speech. In *Proceedings of Conf. on Empirical Methods in Natural Language Processing*.

M. Ostendorf. forthcoming. Prosodic boundary detection. In M. Horne, editor, *Prosody: Theory and Experiment. Studies Presented to Gosta Bruce*. Kluwer.

B. Pellom, W. Ward, J. Hansen, K. Hacioglu, J. Zhang, X. Yu, and S. Pradhan. 2001. University of Colorado dialog systems for travel and navigation.

J.R. Quinlan. 1992. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Robert E. Schapire and Yoram Singer. 2000. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2–3):135–168.

E. Shriberg, A. Stolcke, and D. Baron. 2001. Can prosody aid the automatic processing of multi-party meetings? evidence from predicting punctuation, disfluencies, and overlapping speech. In *Proc. of ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*.

N. Ward and W. Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 23:1177–1207.

Yi Xu. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics*, 25:62–83.