Combined Use of Speaker- and Tone-Normalized Pitch Reset with Pause Duration for Automatic Story Segmentation in Mandarin Broadcast News

Lei Xie, Chuan Liu and Helen Meng
Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Hong Kong SAR of China
{lxie, cliu3, hmmeng}se.cuhk.edu.hk

Abstract

This paper investigates the combined use of pause duration and pitch reset for automatic story segmentation in Mandarin broadcast news. Analysis shows that story boundaries cannot be clearly discriminated from utterance boundaries by speaker-normalized pitch reset due to its large variations across different syllable tone pairs. Instead, speaker- and tone-normalized pitch reset can provide a clear separation between utterance and story boundaries. Experiments using decision trees for story boundary detection reinforce that raw and speaker-normalized pitch resets are not effective for Mandarin Chinese story segmentation. Speaker- and tone-normalized pitch reset is a good story boundary indicator. When it is combined with pause duration, a high *F*-measure of 86.7% is achieved. Analysis of the decision tree uncovered four major heuristics that show how speakers jointly utilize pause duration and pitch reset to separate speech into stories.

1 Introduction

Pitch reset refers to the speaker's general pitch declination through the course of a speech unit, followed by a reset to a high pitch at the start of next speech unit, as shown in Figure 1(a). The speech unit may be of different levels of granularity (Tseng et. al., 2005), such as a speech segment that conveys a central topic (e.g. a news story), a prosodic phrase group (PG) or an utterance. These units are often separated by pauses. Pauses and pitch resets were shown to be effective story boundary indicators in English broadcast news segmentation (Shriberg et. al., 2000; Tür et. al., 2001). These previous efforts specifically point out that pause durations are longer and pitch resets are more pronounced at story boundaries, when compared to utterance boundaries in English broadcast news. However, such story segmentation approaches may be different for a tonal language such as Mandarin Chinese. The use of similar prosodic features for Chinese news story segmentation deserves further investigation. The main reason is that Chinese tonal syllables may complicate the expressions of pitch resets. Chinese syllable tones are expressed acoustically in pitch trajectories, i.e, different tones show different pitch value ranges and trajectory patterns,[1] as shown in Figure 1(b). Initial work in (Levow, 2004) has shown that Mandarin words at story ending positions show a lower pitch as compared with words at non-story-ending positions. In this paper, we present a data-oriented study to investigate how the tonality of Mandarin syllables affects pitch resets at utterance and story boundaries. To alleviate the effects from tonality, we propose to use speaker- and tone-normalized pitch reset with pause duration to separate Mandarin broadcast audio stream into distinct news stories.

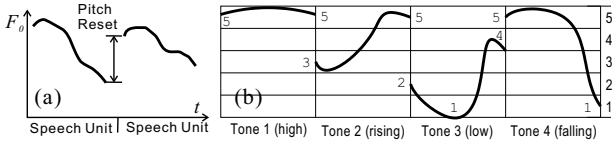

Figure 1: (a) Pitch reset phenomenon between speech units; (b) Pitch trajectories for the four Mandarin basic syllable tones. The speaker pitch range is segmented to five zones from high to low. The pitch trajectories of the four tones are 5-5, 3-5, 2-1-4 and 5-1, respectively.

2 Task and Corpus

In a continuous audio stream of broadcast news, there are programs that consist of speaker changes among anchors, reporters and interviewees. Other programs may contain a sequence of news stories reported by a single speaker. We focus on the latter kind in this investigation, because the combined use of pause duration and pitch reset to punctuate the end of a story and the beginning of the next carries many speaker-dependent characteristics.

We select a subset of TDT2 VOA Mandarin broadcast news corpus (LDC, 1998) and manually extract the news sessions reported by a single speaker. We also annotate

[1] http://www.mandarinbook.net/pronunciation/

Proceedings of NAACL HLT 2007, Companion Volume, pages 193–196,
Rochester, NY, April 2007.

Table 1: The TDT2 subset used in this study.

Nature	Mandarin news sessions reported by a single speaker (13.4 hours)
# of News Sessions	175 (Training: 74, Development: 50, Testing: 51)
Mean Session Duration	276 seconds, 1071 Mandarin characters
# of Story Boundaries	1085 (Training: 442, Development: 316, Testing: 327)
# of Speakers	11 (7 females and 4 males)
Mean Story Duration	36 seconds, 105 Mandarin characters
Transcriptions	Dragon ASR recognizer, GB-encoded word-level transcriptions in XML format

the news story boundaries in this subset. These single-speaker sessions typically contain between 3 to 9 short news stories separated by pauses and constitute about 30% of the entire TDT2 Mandarin corpus (by time duration). The selected subset is divided into training, development and testing sets. Details are shown in Table 1.

3 Region of Interest and Pitch Extraction

Previous work on English news segmentation (Shriberg et. al., 2000) measured pitch resets at inter-word boundaries. Since Chinese news transcripts come as a character stream and each character is pronounced as a tonal syllable, it is more reasonable to investigate the pitch reset phenomenon at the syllable level. We assume that a story boundary must occur at an utterance boundary. The utterances are separated by labeled pauses in the VOA transcriptions ([P] in Figure 2) and a story may contain various utterances (between 2 to 38 in the corpus). Therefore, we only investigate pitch resets in inter-syllable regions across two consecutive utterances as shown in Figure 2. This is reasonable because there are only 6 story boundaries (out of 1085) that are not signaled by pause breaks in the corpus. The region of interest (ROI) is limited to only two tonal syllables, i.e., the last tonal syllable of the previous utterance and the first tonal syllable of the following utterance. We have performed experiments on window length selection and results have shown a wider window does not bring a noticeable improvement.

Raw pitch values are extracted by the YIN pitch tracker (Cheveigné et. al., 2002). The output pitch trajectories are ranked as "good" and "best" by the pitch tracker. Pitch values for unvoiced and pause segments are assigned to be zero. We keep the "best" pitch trajectories for pitch reset measurements. We focus on pitch resets in the ROIs and thus obtain pitch contours for the left and right tonal syllables for each ROIs. However, the corpus transcription does not provide time annotations for those tonal syllables. Therefore, in the pitch trajectory of an audio stream, we search forwards and backwards on both sides of the pause segment for the nearest non-zero pitch measurement sequences. The two pitch sequences found are used as the pitch contours for the left and right tonal syllables of the ROI, respectively. This approximation is reasonable because a Mandarin tonal syllable usually exhibits a continuous pitch contour within its time duration.

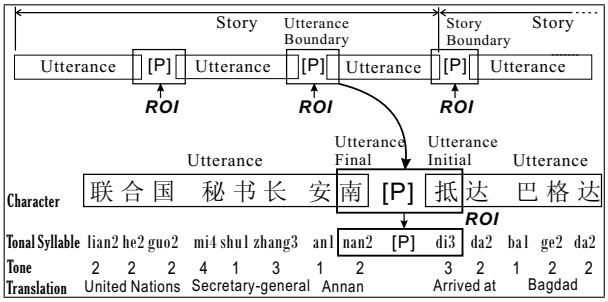

Figure 2: Region of interest(ROI) for pitch reset measure.

4 Speaker- and Tone-Normalized Pitch Reset Analysis in Mandarin Broadcast News

We investigate the pitch reset behavior in the ROIs, i.e., the pitch jump between the left and right tonal syllables at utterance and story boundaries across all corpus audio. Since pitch is a speaker-related feature, we adopt speaker-normalized pitch reset, defined as

$$PR = F_{0r} - F_{0l}, \quad (1)$$

where F_{0l} and F_{0r} are the speaker-normalized pitch for the left and right tonal syllables in the ROIs, which are calculated using

$$F_0 = (f_0 - \mu^s_{f_0})/\sigma^s_{f_0}. \quad (2)$$

f_0 denotes the mean value of the pitch contour of a tonal syllable uttered by speaker s. $\mu^s_{f_0}$ and $\sigma^s_{f_0}$ are the pitch mean and standard deviation calculated for speaker s over all the ROIs of speaker s in the corpus.

We measure the speaker-normalized pitch resets in all ROIs, and categorize them into two boundary types, i.e. utterance boundary and story boundary. To show the effects of tonality in pitch movement, we also categorize the pitch resets by different tone combinations (16 combinations for 4 Mandarin tones[2]). Figure 3 plots the mean PR of each tone combinations for the two boundary types calculated on the corpus data. We see that the pitch reset phenomenon holds for all tone combinations, even for the tone pair (1,3) (i.e. high, low) that has a very small reset. We perform *t*-tests ($p < 0.0025$, one-tailed), which show that for a given tone pair across a boundary, there is a significant difference in PR between an utterance boundary and a story boundary. However, the PR values vary greatly across different tone pairs. For example,

[2]The neutral tone is not considered here since its pitch pattern depends heavily on its neighboring tonal syllables.

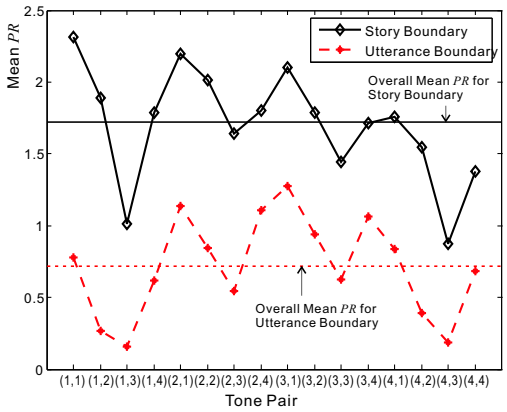

Figure 3: Mean speaker-normalized pitch reset of the 16 tone pairs for story and utterance boundaries.

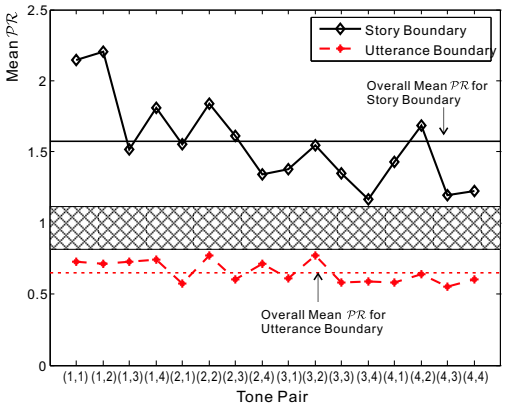

Figure 4: Mean speaker- and tone-normalized pitch reset of the 16 tone pairs for story and utterance boundaries.

pitch resets are reduced for the tone pairs (1,3) and (4,3), but are pronounced for the tone pairs (3,1) and (2,1). The *t*-test ($p < 0.0025$, one-tailed) shows that the PR difference between utterance boundaries and story boundaries are *not* significant. This motivates us to formulate a definition for speaker- and tone-normalized pitch reset.

The speaker- and tone-normalized pitch reset is defined as:

$$\mathcal{PR} = \mathcal{F}_{0r} - \mathcal{F}_{0l}, \quad (3)$$

where $\mathcal{F}_{0l}$ and $\mathcal{F}_{0r}$ are the speaker- & tone-normalized pitch for the left and right tonal syllables in the ROIs, respectively, defined as

$$\mathcal{F}_0 = (F_0 - \mu^{\tau}_{F_0})/\sigma^{\tau}_{F_0}, \quad (4)$$

where F_0 is the speaker-normalized pitch in Equation (2) of a tonal syllable with tone τ. $\mu^{\tau}_{F_0}$ and $\sigma^{\tau}_{F_0}$ are the pitch mean and standard deviation calculated for the tonal syllables with tone τ over all ROIs in the corpus. Figure 4 plots the mean $\mathcal{PR}$ of each tone combinations for the two boundary types calculated on the corpus data.

Figure 4 shows a clear separation in speaker- and tone-normalized pitch reset ($\mathcal{PR}$) between utterance and story boundaries (shade area in Figure 4). This result is statistically significant based on a *t*-test ($p < 0.0025$, one-tailed). This observation suggests that speaker- and tone-normalized pitch reset may be an effective story boundary indicator for Mandarin broadcast news.

5 Experiments on Story Boundary Detection

We perform experiments on story boundary detection at the ROIs in the corpus. Since all ROIs are utterance boundaries, of which only some are story boundaries, we take a "*hypothesize and classify*" approach in order to strike a good balance between recall and precision. We first hypothesize the occurrence of a story boundary if the ROI has a pause duration that exceeds a threshold. This is followed by a decision tree classifier that decides on the existence of a story boundary. We used Quinlan's C4.5-style decision tree (Quinlan, 1992) as the classifier, implemented by the IND toolkit.[3] The pause duration threshold was selected by a heuristic search procedure described as follows: We experimented with pause durations ranging from 0.1 to 4 seconds with step size of 0.1 second. In each case, we hypothesized raw boundaries in the training and development sets. A decision tree was then grown using the raw boundary hypotheses of the training set, and tested on the raw boundary hypotheses of the development set. The pause duration leading to the highest F-measure on the development set was selected as the optimal threshold for the further experiments on the testing set.

We develop seven story boundary detectors according to the features used (see Table 2). The boundary detection results on the testing set are shown in Table 2. From Table 2, we can see that the detector using pause duration achieves a high F-measure of 82.2%. This result is reasonable since VOA Mandarin news broadcast makes large use of long pauses at story boundaries, especially at news sessions reported by a single speaker. The detector using raw pitch reset ($pr = f_{0r} - f_{0l}$) only gets a F-measure of 50.8% and the speaker-normalized pitch reset (PR) achieves a slightly better F-measure of 55.3%. Speaker- and tone-normalized pitch reset ($\mathcal{PR}$) achieves a superior performance with an F-measure of 71.1%. This result is consistent with the observations in Section 4. The story boundary indicative ability of speaker-normalized pitch reset is affected by the tonality of Mandarin syllable. Speaker- and tone-normalized pitch reset can alleviate the effects, thus leading to a better discrimination. Based on Table 2, when pause is combined with raw pitch reset, the F-measure degrades from 82.2% to 68.3%. The F-measure reaches 77.4% when we combine pause with speaker-normalized pitch reset. When pause is combined with speaker- and tone-normalized pitch reset (*Pause*+$\mathcal{PR}$), the best F-measure is achieved at 86.7%.

[3] http://ic.arc.nasa.gov/projects/bayes-group/ind/

Table 2: Story boundary detection experiment results(%)

Feature	Recall	Precision	F-Measure
Pause	77.1	88.1	82.2
pr	52.0	49.7	50.8
PR	56.6	54.1	55.3
$\mathcal{PR}$	70.3	72.0	71.1
Pause+pr	66.4	70.3	68.3
Pause+PR	72.2	83.5	77.4
Pause+$\mathcal{PR}$	82.6	91.3	86.7

Table 3: Heuristics for story boundary decision

No.	Description	Story Boundary?
1	Pause duration is short ($P < 1.475$) and pitch reset is small ($\mathcal{PR} < 0.401$)	No
2	Pause duration is short ($P < 1.475$) and pitch reset is huge ($\mathcal{PR} > 1.112$)	Yes
3	Pause duration is long ($2.315 \leq P < 4.915$) and pitch reset is big ($\mathcal{PR} > 0.715$)	Yes
4	Pause duration is long ($P \geq 4.915$) and pitch reset is low ($\mathcal{PR} < 0.3513$)	No

Figure 5 shows the top levels of the decision tree obtained using the *Pause*+$\mathcal{PR}$ set. We can observe the complementarity between pause duration and pitch reset in story boundary detection. This may be summarized in terms of four major *heuristics* shown on the tree (labeled as 1 to 4 in Figure 5). These heuristics cover about 83% decisions made on the testing set, as described in Table 3.

Heuristics 2 is mainly used to detect possibly missing story boundaries with short pauses caused by speaker speaking style, e.g., reporters Li Weiqing and Yang Chen tend to use short pauses to separate news stories, but they tend to offset the reduced pauses with pronounced pitch resets to signify story boundaries. Heuristics 4 detects possibly false alarms due to broadcast interruptions in boundary detection. These interruptions (i.e. silences) usually occur within a news story and may last for several seconds (usually > 5 $seconds$).

6 Summary and Future Work

This paper investigated the combined use of pause duration and pitch reset for automatic story segmentation in Mandarin broadcast news. Pitch reset analysis on Mandarin broadcast news shows that story boundaries cannot be discriminated from utterance boundaries by speaker-normalized pitch reset, because speaker-normalized pitch reset varies greatly across different tone pairs of boundary syllables. This motivates us to investigate the speaker- and tone-normalized pitch reset. Analysis shows that speaker- and tone-normalized pitch reset can clearly separate utterance boundaries from story boundaries across all tone pairs. This result shows the difference between English and Chinese. Previous work for English (Shriberg et. al., 2000; Tür et. al., 2001) shows that speaker-normalized pitch reset is effective. This work shows that the same measurement is not sufficient for Chinese; instead we need to use speaker- and tone-normalized pitch reset in Chinese story segmentation. When pause duration is combined with speaker- and tone-normalized pitch reset, the best performance is achieved with a high F-measure of 86.7%. Analysis of the decision tree uncovered four major heuristics that show how speakers jointly utilize pause and pitch reset to separate speech into stories.

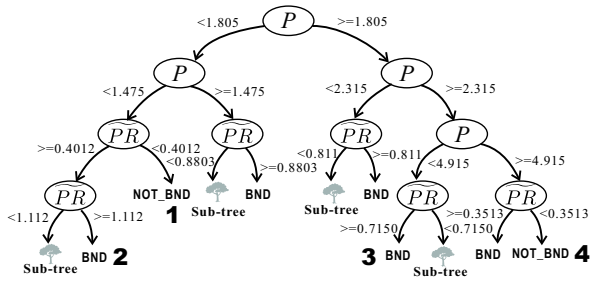

Figure 5: Decision tree for story boundary classification based on the *Pause*+$\mathcal{PR}$ feature set. BND denotes story boundary, and NOT_BND denotes not story boundary.

Future work will investigate the pitch reset phenomenon in Cantonese broadcast news, because Cantonese is another major Chinese dialect with more complicated tonal characteristics. We also plan to incorporate prosodic cues with lexical cues to further improve performance in Chinese story segmentation.

References

Shriberg E., Stolcke A., Hakkani-Tür D. and Tür G. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Comm.*, 32(1-2):127–154.

Tür G. and Hakkani-Tür D. 2001. Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation. *Computational Linguistics*, 27(1):31–57.

Levow G. A. 2004. Prosody-based Topic Segmentation for Mandarin Broadcast News. *Proc. of HLT-NAACL*, 137–140.

The Linguistic Data Consortium. 1998. *http://projects.ldc.upenn.edu/TDT2/.*

de Cheveigné A. and Kawahara H. 2002. Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustic Society of America*, 111(4):1917–1930.

Tseng C. Y., Pin S. H., Lee Y., Wang H. M. and Chen Y. C. 2005. Fluent speech prosody: Framework and modeling. *Speech Comm.*, 46:284–309.

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