Analyzing input source in the acquisition of Mandarin Tone2 and Tone3: A Smoothing Spline Anova Approach to Tones in Adult and Child-directed Speech

Shuo Zhang

December 27, 2013

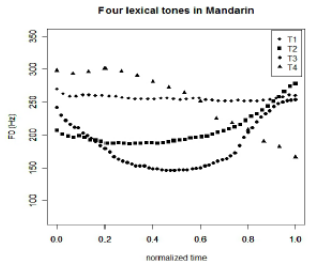
Part I
Introduction

In the research of Mandarin tone acquisition, Tone 2 (low rise) and Tone 3 (contour dipping tone) have been shown to be the most difficult to acquire in both L1 and L2 acquisition (Li&Thompson 1978, Jongman et al 2002,Chang 2011, Guo 2008), usually to be acquired at a later stage (in the case of L1) and/or with greater difficulty in production and perception (mostly in L2, Lai and Zhang 2008).

Many perceptual experiments have been conducted to show the contribution of relevant acoustic cues in T2/T3 distinction among native or non-native speakers. These acoustic characteristics include the timing of the truning point (Shen & Lin 1990, Moore & Jongman 1997), the depth of F0 difference in the initial fall (Shen et al 1990,1991,1993), syllable duration (Blicher et al. 1990, Chang 2011), amplitude (Chuang et al 1997), voice quality (T3 is associated with creaky voice, Yang 2011), and the isolation point in a gating paradigm (Lai and Zhang 2008). Figure 1 shows a scheme of the tone contours of the four tones in Mandarin in citation form.

In particular, the durational cue in T2/T3 distinction has received continuous attention. Early literature (Lin 1965) noted that the duration of T3 is consistently the longest when produced in isolation among all four tones, with T4 being the shortest. Shen (1990) confirmed that T3 is longer than T2 in isolation form in an acoustic study. Blicher et al (1990) conducted perception experiment and concluded that lengthening constitutes a positve perceptual cue in favor of T3 (in isolation). Chang (2011) extended a perceptual experiment by artificially time-normlizing T3 to the shorter duration of T2 in isolation form. They found that while L2 speakers' performance suffered significantly in terms of tone identification accuracy and reaction time, L1 speakers only showed longer reaction time. They concluded that duration cues may constitute a facilitation cue for L1 speakers, whereas in the case of

Figure 1: Mandarin tones in citation form



L2 speakers it is a crucial cue.

While these studies found eveidence for duration (and other acoustic features) as important cues for *identifying* T2/T3 in isolation, two weaknesses challenge the aforementioned studies in their validity in establishing durational cue as a crucial factor to explain the difficulty in *acquiring* T2 and T3. First, overall, the body of literature mentioned above test and analyze T2/T3 perception in isolation. As a basic *a priori* observation, there exists a significant difference between actual realization of tones in connected speech in normal speech rate, and tones produced in isolation/citation form in an experiment (this is recently quantitatively and experimentally described by Xu (2002)'s Target Approximation Model of tone production). This difference exists in many dimensions of acoustic characteristics, and most notably in duration and contour shape.[1] Second, as already pointed out by Shih(2007), duration cues cannot be reliably shown to be crucial for T2/T3 distinction due to the observation that in running speech, T3 often becomes the shortest among all four tones since its rising tail is not realized (becoming a low dipping or low level tone). It is therefore not appropriate to use isolation forms to explain the difficulty of T2/T3 acquisition, where connected speech is the main source of input (this is true for both L1 and L2 acquisition, with a possible exception of early stages of L2 acquisition in classrom). For instance, we might want to reconsider Chang(2011)'s experiment and investigate the consequences if we assume T3 to be consistently shorter than T2.

The current project seeks to take a preliminary step toward understanding the acoustic properties of tones (with special focus on T2/T3) in connected speech, by taking the assumption that connected speech is a primary input source to language acquisition. In this pilot study, I analyze recorded speech data from Mandarin native speakers from two sources: the first data set consists of three adult speakers in a semi-spontaneous production task in a casual speech style. The second data set contains recorded child-directed speech from two adult speakers (drawn from the CHILDES database). A total number of 348+327 tone contour units in these two dataset are extracted with Praat and analyzed using the Smoothing Spline ANOVA (SSANOVA) model in R. In this analysis, I focus on these two dataset to explore several preliminary questions in the input source (i.e., connected speech) of tone acquisition: (1) what is the duration difference between T2 and T3? (2) What is the overall shape of tonal contours realized in connected speech in all four tones? (3) Can these give a possible explanation to why T2/T3 is difficult to acquire? (4) Is there a qualitative and/or quantitatve difference represented in the current datasets regarding tones in adult and child-directed speech? If so, what is the implication?

Part II
Methodology

2.1 The Data Sets

The current paper analyzes data from two input sources in language acquisition: adult and child-directed speech. For L1 child, these two sets of data serve as a naive (or ideal) model of representation for the types of input speech that they are exposed to. The distribution and contribution of these two types in the L1 input is unclear, and may depend on individual cases. For L2 speakers, their only input is adult speech, yet in earlier stages of classroom instruction, they may be exposed to more careful and clearly articulated speech in language pedagogy. These are not represented here. Nonetheless, the choice of these two data sets reflects an attempt to represent the type of connected speech in language acquisition, and it is the goal of our model to account for the differences/similarities between them, and how they are related to the T2/T3 acquisition.

[1] Even though I challenge this methodology, however, this paradigm of using tones in isolation form in experiments is understandable, partly due to the challenge that in connected speech, word identification does not depend on tone identification alone. Xu and Patel (2010) has shown that Mandarin speech is 90% intelligible in a non-noisy background while played in monotone. Therefore, the challenge of solely investigating tone perception using connected speech is to eliminate the probablistic contextual and segmental cues in the speech that may give away the answer.

The adult data set is drawn from an existing data set composed of 11 young adult speakers of Mandarin in Beijing performing a semi-spontaneous speech production task in a conversational and casual speech style. The participants are given a list of words and to produce sentences in the format of:

(1)wo3 xi3huan1/bu4 xi3huan1 ___.

"I like/ don't like ___".

The participants pick items in the word list based on whether they like it or not, and produced 98 sentences per person. The current data set is a subset of the original data. A total of 348 tone contour units, evenly distributed across all four tones, are segmented and extracted. The format of this task determines that this dataset is a more restricted representation of the entire universe of adult speech that the L1 child is exposed to, especially for our purposes, considering the vast amount of possible variations that exists in the tone contours of adult speech in more varied positions, coarticulations, and emotional states in many dimensions of the acoustic space. The consequence of this is that the results of the currect tone contour data analysis is more restricted and "clean" (if you will) than the actual adult speech data, which presents a greater challenge for the learner.

The child-directed speech data set (Child dataset henceforce) is from the CHILDES database[2]. The two speakers selected are caregivers telling a story of little frog to a child. In each case, the child is between two and three years old, which makes it possible for the story to be understood. The storyteller later asks the child to reiterate the story to her and the child's parents by asking the child questions and giving hints as reminders to how the story progresses. That part of the recording is not included in the analysis.

2.2 Data Preprocessing

In this step I segment the recorded speech data to identify and label tone units, and extract the tone contour units to produce a csv file for statistical analysis. To do this, several praat scripts are developed to automate the process to the extent possible. Here is a short description for the scripts and procedures developed for the data preprocessing:

1. `intensity-seg.praat/voicing-seg.praat` and hand-labelling of tones:this set of script is used to automate segmentation to a maximum degree possible. The intensity segmentation first creates a intensity contour object inside Praat on the utterance/word/phrase level, segmenting sound from silence (including inter-word silence but in this case does not include silence of a consonant closure). For each sounded segment, a second level of intensity-based or voicing based[3] segmentation takes place, where the valleys in intensity contour are located and labeled. These usually correspond to syllable boundaries, while in many cases it can result in more fine-grained segments than desired. After the segmentation is done, each syllable in the textgrid is then hand labeled according to their tone category (1,2,3, or 4). In this process, automatically obtained syllable boundaries are adjusted wherever applicable. If an interval does not contain a usable tone contour, it is simply left blank.Due to the continuous nature of vocal folds and the produced pitch contour, the tone units are segmented in a way that it only includes the core part of the syllable where the voicing is reliable[4]. The intervocalic part of the pitch contour, especially when they cross word boundary, is to be excluded, as their contours mostly consist of rises and falls connecting the adjacent tones. Nonetheless, a great number of tones still show a considerable 'detour' that connects the current pitch target to the next, a phenomenon noted in Xu's Target Approximation Model (2002) to be predominant in normal and fast speech rate.

[2]The recording of this file is located at http://childes.psy.cmu.edu/media/EastAsian/Chinese/Beijing2/F2/ (accessable from 'media' on homepage), whereas a CHILDES-formatted transcript can be found in the 'database' following the EastAsian/Chinese/Beijing2/F2 path.

[3]Here it turns out that intensity based segmentation better facilitates the hand labelling.

[4]Consistent labelling practice is attempted in order to have a meaningful comparsion of tone durations.

2. `get-tones.praat`:This script simply iterates through all non-empty intervals in the tone tier and find out how many tones have been labeled for each tone category.

3. `extract-contour.praat`: This is the most crucial tool in the extraction of pitch contours to automatically produce a `csv` file. It generates a pitch object in Praat, in which pitch is computed using autocorrelation algorithm (pitch step 0.01s with pitch floor = 55Hz and pitch ceiling = 700Hz for Child dataset, due to the expressive big pitch range of the female speaker's child-directed speech, and standard setting for Adult dataset) for each frame of the pitch contour unit(i.e., PCU, a unit of one tone contour, usually corresponding to a syllable) that are identified and labeled earlier. Crucially, each PCU is assigned a pitch_con ID so that each of the pitch values consisting of this particular contour is identified collectively by this ID, facilitating the statistical analysis in R. This script writes all values into the columns of the final `csv` file, including time, pitch, pitch_con (PCU_ID), tone (tone category of the current PCU), and duration (duration of the current PCU). Speaker information is encoded into PCU_ID so that each PCU_ID in the entire data set (Adult and Child) is unique.

2.3 Statistical Analysis and SSANOVA Modeling

The statistical analysis on duration and pitch height are carried out in statistical computing tool R. To model the average shape of the tone contours with a Bayesian confidence interval, the Smoothing Spline ANOVA modeling is carried out using the `gss` package implemented in R (Gu 2002).

Smoothing splines are a type of natural cubic spline, which is a piecewise polynomial function that connects discrete data points called knots. Smoothing splines include a smoothing parameter to find the best fit when the data tend to be noisy. More specifically, the function defining the smoothing spline contains two terms: one that attempts to fit the data and one that penalizes a fit which does not have the appropriate amount of smoothness.The smoothing spline is estimated by minimizing the Following function:

$$G(x) = \frac{1}{n}\sum_{\text{all } i} (y_i - f(x_i))^2 + \lambda \int_a^b (f''(u))^2 \, du, \tag{2}$$

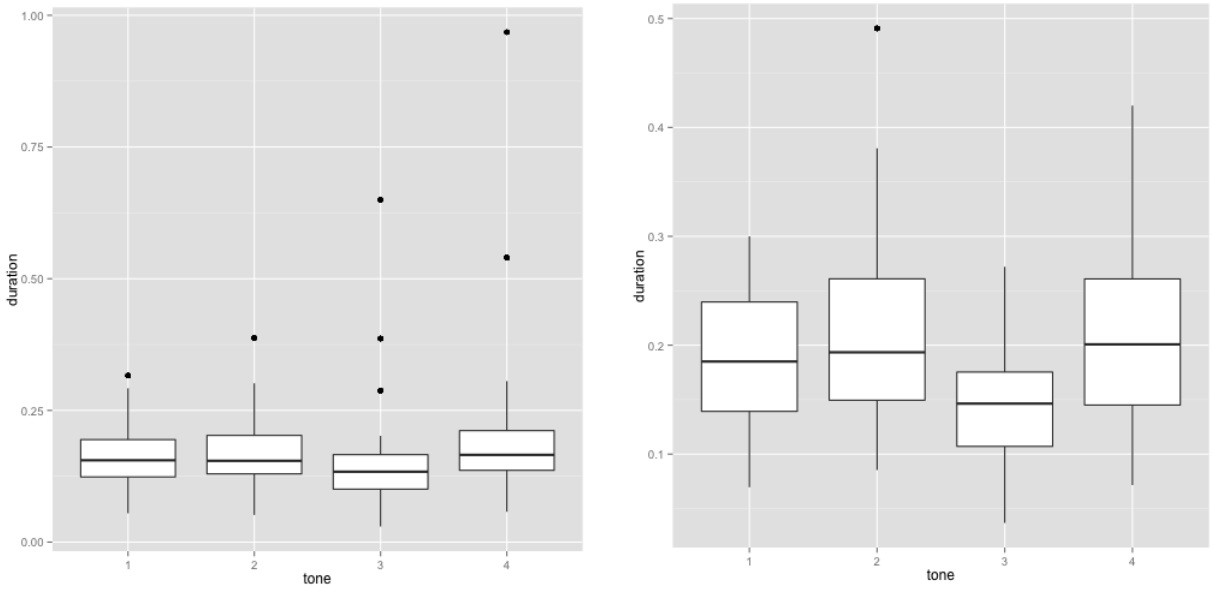
where n is the number of data points, λ is the smoothing parameter, and a and b are the x coordinates of the endpoint of the spline. (Davidson 2006)

The SS ANOVA has been used in applications that require a statistical technique to determine whether the shapes of multiple curves are significantly different from one another. This technique has been used to compare tongue trajectories in ultrasound imaging data for the articulatory phonetics (Davidson 2006). It is used to model contour shapes (analogous to the linear model in individual data points) in a series, trajectories, or time series of data points. The SS ANOVA model is of the following form. Each component of f is estimated with a smoothing spline:

$$f = \mu + \beta x + \text{main group effect} + \text{smooth}(x) + \text{smooth}(x;\text{group}). \tag{3}$$

In the current analysis, SSANOVA is used to model the average contour shapes of all the PCUs of the same tone category. If there are less variants in the contour shapes, the shapes can be represented by a single contour. If the contour shapes of all PCUs belonging to the same tone category are vastly different, then it may be represented by a group of average contour shapes. Since analysis on duration and pitch height are already carried out using traditional statistical analysis in R, in the analysis of contour shapes, for the ease of comprison, pitch contours are time normalized so that all contours stretch from 0 to 2 second while preserving its contour shape. They are also normalized in pitch height so that all pitch values for a data set (Adult or Child) have a mean pitch of 0 and standard deviation

Figure 2: Duration difference of tone categories (a) Adult(left) (b) Child data sets



of 1. For each tone category in a data set, SSANOVA computes the 'average' contour(s) of the PCUs by building a model, and it computes the Bayesian confidence interval by generating predicted data from the model and computes the error estimate.

Part III
Results and Discussion

3.1 Duration

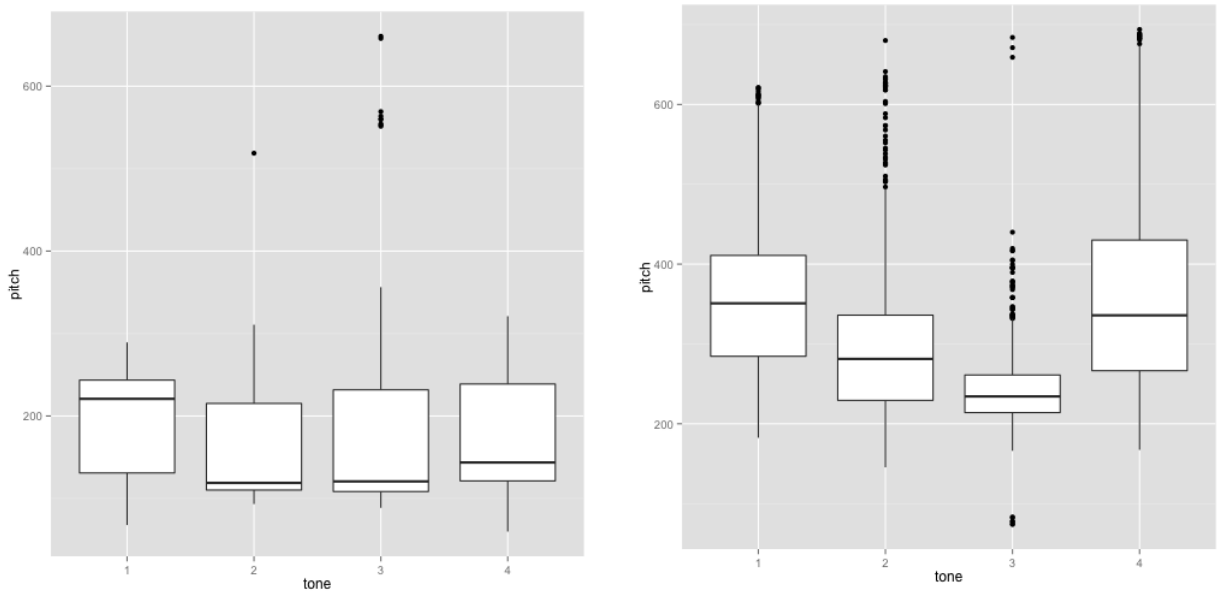
The results of durational differences across four tones in both the Adult and Child data set are shown in Figure 2 (a) and (b). The plots show that by all measures, tone 3 has shorter durations than tone 2. In fact, this confirms Shih(2007)'s observation that Tone 3 is often the shortest among all four tones in connected speech. Comparing Adult data with Child data, we do not see drastic differences between the two (the adult data is visually compressed in this graph because it has quite a few outliers, but otherwise they're actually in the same range as the Child data), with Child data having longer duration by 0.02s in a two-sample t-test ($t = -11.219$, $df = 7521.216$, p-value < 0.001). This is due to the slower speech rate in child-directed speech.

In both cases, T3 is shorter than T2. A linear regression model (ANOVA) shows that T3 is significantly shorter than T2 ($p<0.001$ in both data sets) in both adult speech and in child-directed speech. This supports the argument (Shih 2007) that the assumption that T3 has longer duration than T2 does not explain the role of duration in T2/T3 acquisition.

3.2 Pitch Height

Figure 3 shows that the pitch range, especially on the upper range is a lot higher in the Child data set than the Adult. This is due to a number of reasons. First, the adult dataset consists of one male speaker and two female speakers. The male speaker contributes to lower overal pitch range. The

Figure 3: Pitch height in (1)Adult(left) (2)Child data sets



two speakers in the Child dataset are both female with higher voice. Second, in the Child dataset, we observe that the highly expressive prosody in the child-directed speech produced many outliers in pitch range even for these two female speakers who have already quite high pitch register.

On the other hand, we observe not a big difference in Adult data between the pitch height of T2 and T3. In fact, a linear regression model (ANOVA) shows that pitch overall in T3 is slightly higher than T2 with a 17Hz difference ($F(1795,1)=31.57, p<0.001$). This is not expected given the general impression that T3 is low and flat in connected speech. However, to foreshadow our results in SSANOVA, this result is not unexpected given the shape of the T3 contour in this data set. Due to the short duration of T3, the average shape of T3 is highly distorted from the citation form, showing a long falling tail on the left side, prior to its low flat pitch target. This shows the prediction of Target Approximation Model (Xu 2002): when the pitch target is short and fast, we observe that a large portion of the syllable bears the connecting tonal contour sliding down from the previous syllable pitch target, and the true target of the current syllable (T3) is barely reached before the speaker is moving on to the next target. In the child data set, however, this effect is overshadowed by the big pitch range in the story-telling speech (in which T2 has overall higher pitch). Therefore, in the Child dataset, our model shows that T3 is significantly lower in pitch than T2 ($F(1776,1)=206.5$, $p<0.001$).

3.3 Pitch Contour Shapes for T2 and T3: SSANOVA Models

The average pitch contour shapes computed by SSANOVA is shown in Figure 4-7, comparing the two datasets (dashed line shows the Bayesian confidence interval). The original raw contours constitute a more noisy plot showing pitch moving in all directions. However, by computing the contribution of each PCU to an average shape, the SSANOVA has produced cleaner contours representing the most typical and strong factors in this group. One exception to this is T1, which is shown to have more variations than other tones. This is somewhat expected since T1 is flat in citation form, and any other pitch movement (connecting pitch targets) may distort the shape of T1 in various ways.

To compare T2 with T3, we observe that the T2 shape is the most clean and closest to its cannonical rising intonation in citation form. T3, on the other hand, turned out to have a rather unexpected shape that resembles T4. This has been discussed earlier: it is possible that T3, as the shortest tone, is the

Figure 4: Adult(left) and Child: Tone 1 contours

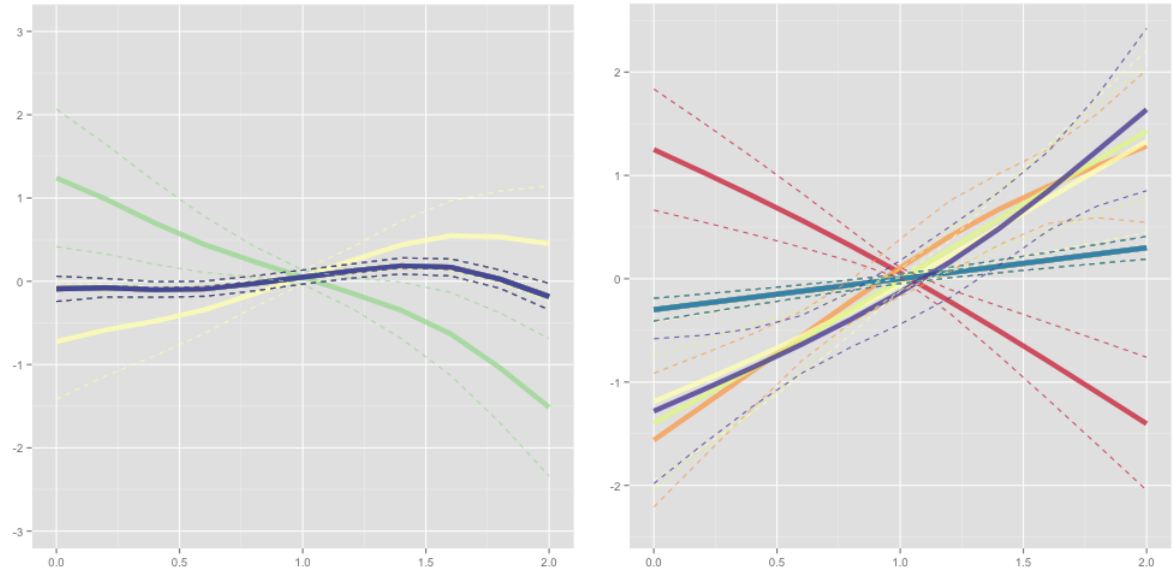


Figure 5: Adult(left) and Child: Tone 2 contours

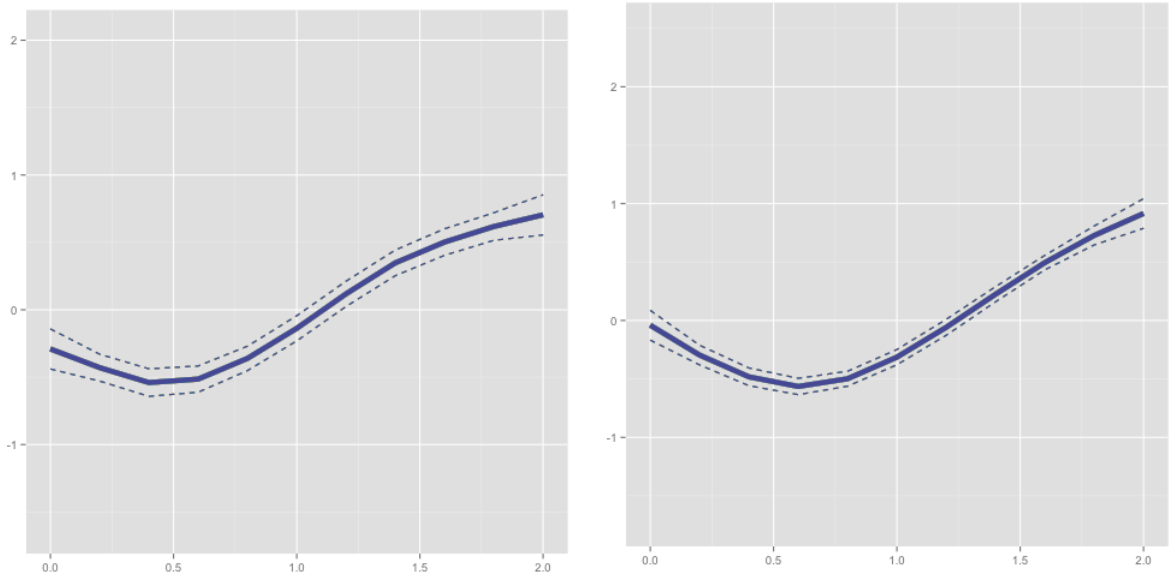


Figure 6: Adult(left) and Child: Tone 3 contours

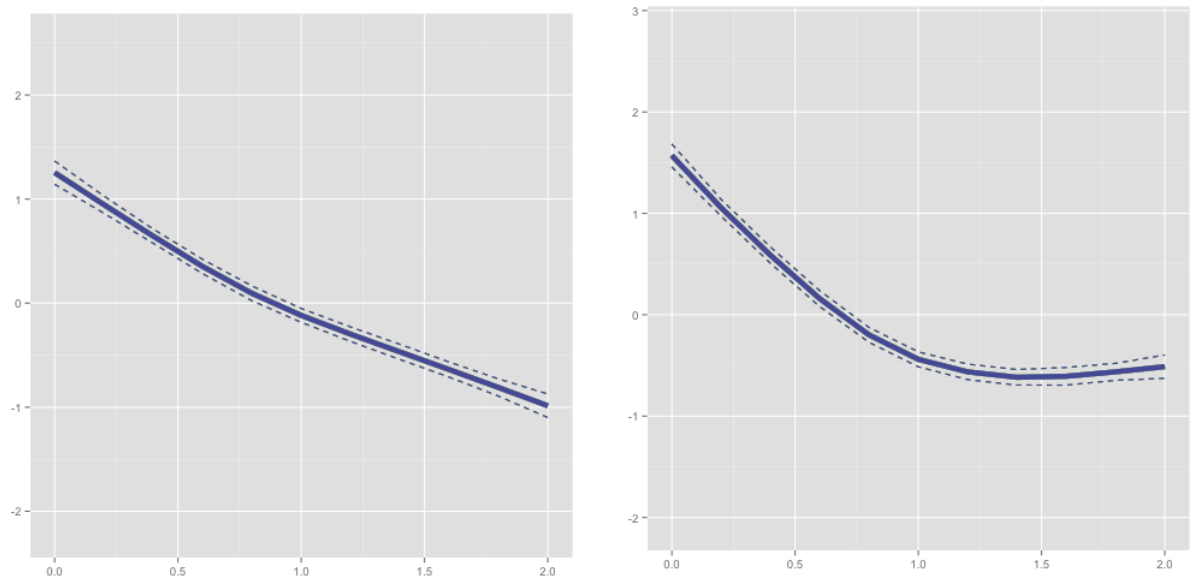
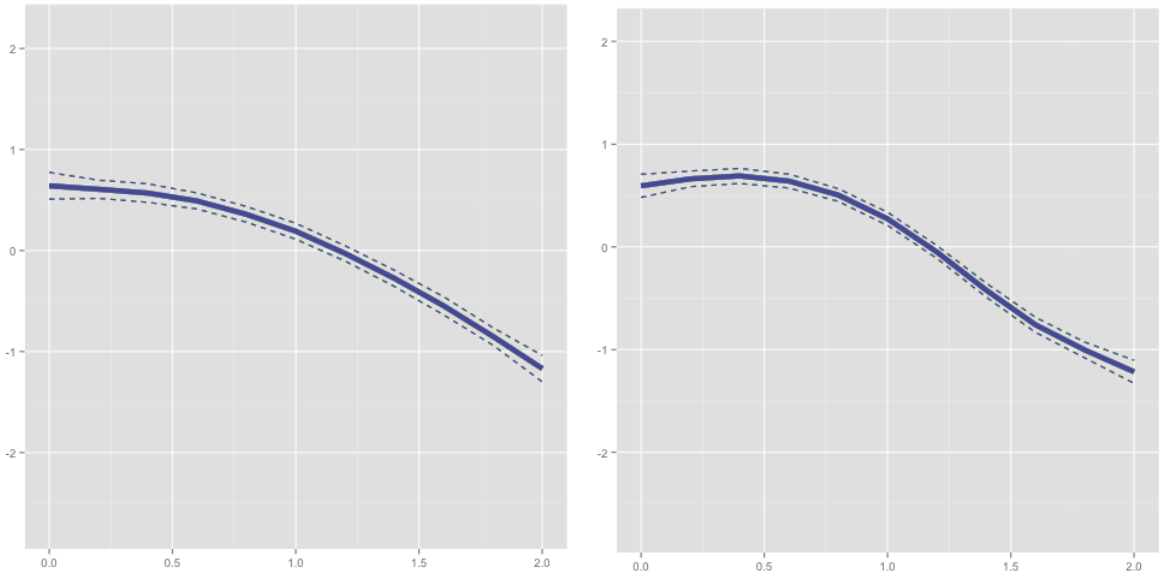


Figure 7: Adult(left) and Child: Tone 4 contours



most susceptible to distortion by the inter-tonal pitch movement. This can be further confirmed by examining the shapes of T3 and T4: While T3 has a sharp fall and (more obvious in the Child dataset) a flat phase in the later part of the tone (which is its true target), the T4 has a flat phase in the initial stage of the contour and then the fall becomes sharper. Therefore, their shapes resemble each other because of entirely different mechanisms.

In this case, however, we observe that T2 and T3 are not confusable at all: they have totally different contour shapes in connected speech, which is expected from the observations made *a priori* (i.e., as pointed out earlier, it is commonly known that T3 is realized as a short dip tone without the rising tail in connected speech, Shih 2007). Perhaps the only similarity we can draw between T2 and T3 in terms of contour shapes are that they are the only two tones in the four tones to have the concave contour (pointing down) in the same direction. Whether that is perceptually crucial is yet to be decided. Finally, our results show that in the current data sets, there is no significant differences between adult and child-directed speech in their tonal contour shapes.

Part IV
Conclusion

This paper is a pilot study to use connected speech from a small corpus of adult and child-directed speech data set to analyze the acoustic properties of Mandarin tones as the input source for tone acquisition. Results show that with regard to the difficulty in acquiring T2/T3, duration of T3 is signiaifcantly shorter than T2, and the contour shapes of the two tones are significantly different in connected speech. While this result is seemingly unable to provide any explanation to the difficulty in acquiring T2/T3, on the other hand, it does precisely what I have set out to prove: namely, it proves that simply using speech experiments to examining the perception of T2/T3 in isolation form is not sufficient an explanation for the difficulty of T2/T3 acquisition. In particular, the duration cue argued in Chang (2011) is based on the assumption that T2 is much shorter than T3 in isolation, which may explain the perception and tone identification in that experiment (in isolation), but not in how and why acquiring these two tones are difficult in real-time L1 or L2 learners. This analysis threrefore will hopefully serve as a starting point to an extended study on a bigger dataset and an experimental methodology that more seriously takes into account the environment that tone acquisition takes place.

References

[1] Blicher, D. L., Diehl, R. L., & Cohen, L. B. (1990). Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: Evidence of auditory enhancement. Journal of Phonetics, 18(1), 37-49.

[2] Chang, Yung-hsiang Shawn. (2011). "Distinction Between Mandarin Tones 2 and 3 for L1 and L2 Listeners" Proceedings of 23rd NACCL, 1: 84–96.

[3] Chuang, C.-K., & Hiki, S. (1972). Acoustical Features and Perceptual Cues of the Four Tones of Standard Colloquial Chinese. The Journal of the Acoustical Society of America, 52 (1A), 146.

[4] Davidson, Lisa. (2006). Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. Journal of the Acoustic Society of America, 120(1).

[5] Gu, C. (2002). Smoothing Spline ANOVA Models Springer, New York.

[6] Guo, Lijuan. (2008). "Tone Production in Mandarin Chinese By American Students : A Case Study" Proceedings of 20th NACCL,1: 123–138.

[7] Jongman, Allard, and Joan A Sereno. (2002). "L2 Acquisition and Processing of Mandarin Tone." in Li, P et al, Handbook of Chinese Psycholinguistics.

[8] Lai, Yuwen and Jie Zhang (2008). Mandarin lexical tone recognition: the gating paradigm. In Emily Tummons and Stephanie Lux (eds.), Proceedings of the 2007 Mid-America Linguistics Conference, Kansas Working Papers in Linguistics 30. 183-194.

[9] Li, C. N., & Thompson, S. (1977). The acquisition of tone in Mandarin-speaking children. Journal of Child Language, 4(2), 185-199.

[10] Lin, M.-C. (1965). Yingao xianshiqi yu Putonghu shengdiao yingao texing [The pitch indicator and the pitch characteristics of tones in Standard Chinese]. Shengxue Xuebao [Chinese Journal of Acoustica] 2 (1), 8–15.

[11] Moore, C. B., and Jongman, A. (1997) "Speaker normalization in the perception of Mandarin Chinese tones."J. Acoust Soc. Am. 102, 1864-1877.

[12] Shen, X-N. S. (1990). Tonal coarticulation in Mandatin. Journal of Phonetics, 18, 281- 295.

[13] Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. Language and Speech, 34, 145-156.

[14] Shen, X. S., Lin, M., & Yan, J. (1993). F0 turning point as an F0 cue to tonal contrast: A case study of Mandarin tones 2 and 3. Journal of the Acoustical Society of America, 93, 2241.

[15] Shih, C. (2007). Prosody learning and generation. Berlin: Springer.

[16] Xu, Y. and Sun X. (2002). Maximum speed of pitch change and how it may relate to speech. Journal of the Acoustical Society of America 111: 1399-1413.

[17] Xu, Y. and Sun X. (2002). Maximum speed of pitch change and how it may relate to speech. Journal of the Acoustical Society of America 111: 1399-1413.

[18] Yang, R. (2011). The Phonation Factor in the Categorical Perception of Mandarin Tones. In Proceedings of ICPhS XVII.