Tonal Alignment in Different Tone Systems

*Yiqing Zu, Runqiang Yan**

China Research Center, Motorola Labs, Shanghai, P.R. China
*Department of biomedical Engineering, Jiaotong University, P.R. China
Yiqing.Zu@motorola.com

Abstract

This study analyzed pitch distribution on speech corpora of Mandarin Chinese, Cantonese and American English. Each Language or dialect has one male speaker and one female speaker. The results show that the tone language has boarder pitch range in high frequency range for tones realization. As a result, tonal alignment is crucial for tone languages. It seems unnecessary to use boarder pitch range for lexical stress in English.

In Mandarin Chinese, when two tones in a di-syllable word is conflicting and the initial of second syllable is voice, it will introduce a extra pitch peak to the onset of second syllable. When those syllables are used in another context when concatenation is conducted it will sounds un-pleasant perceptively. To investigate tonal alignment, this paper studies the di-syllable prosodic words in a Mandarin TTS speech corpus and concludes that the alignment failure phenomena, which are important to speech technology, have more chances occurring in tone languages.

1. Introduction

The major difference between tonal language and un-tonal language is that tone language has complex tone system in lexical level. The complex tone system means that there may be more than one pitch values in a lexical tone. In tone language, a speaker uses same dimension to realize lexical tone and intonation. Just because there are may be more than one pitch target in a lexical tone, the combination of lexical tone will more complex. Pitch conflicting cases will happen in some tone combinations.

Within a prosodic phrase, there are confusions to determine where is the exactly segmental boundary because segments connect each other with high degree of co-articulation. For example, each Mandarin syllable has a lexical tone, which is aligned to the spectrum of a mandarin syllable. Due to the physical constraint of articulators, as founded by Xu Yi [1,2], extra or delayed pitch peaks are introduced in some tone contexts. We call this phenomenon as alignment failure. The alignment failure is resulted from alignment of different articulators and will cause the discontinuity at concatenation point. All of these will bring un-pleasant feelings when un-appropriate concatenation is conducted. If those phonetic details are processed properly, the smoothness of concatenation point will be improved. In English speech, the pitch peaks in stressed syllable will not delay to the following un-stressed syllable. There is only a high pitch on-set in second syllable. Based on TTS speech corpora, this paper analyzes pitch distributions of tonal and un-tonal languages and will focus the discussion on of tonal alignment problem in these languages.

2. Pitch distributions in different Languages and different dialects

2.1. Pitch distribution in Chinese dialects and English speech data

Using TTS speech corpora of Mandarin female/male, American English male/female and Cantonese female/male speaker, we draw the pitch distributions, as shown in figure 2.1.1a – figure 2.1.1e. The vertical axis is the numbers of pitch samples and the horizontal axis is frequency, scaled from 0 to 600Hz. The scripts of the TTS speech corpora are well designed based on newspapers. The Mandarin corpora cover all of tonal syllables in different sentence position. The English corpora cover variable mono-phone, di-phone, triphone and high frequently used words. The speech signal was recorded by DAT in a studio stored as 16bit, 16KHz sampling rate data. The pitch values are extracted each 10 ms by speech frames of 20 ms using Praat Speech Analysis tool. One pitch sample is related to one frame. Each speech corpus is big enough to describe pitch distribution.

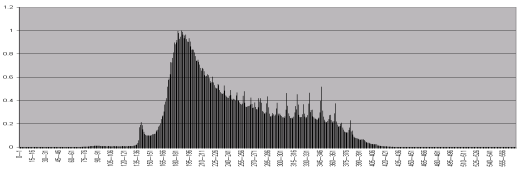

Figure 2.1.1a *Pitch distribution of a Mandarin female speaker*

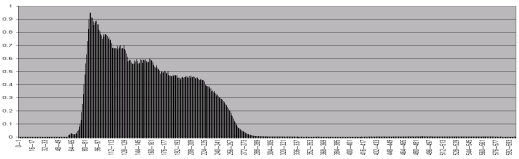

Figure 2.1.1b *Pitch distribution of a Mandarin Chinese male speaker*

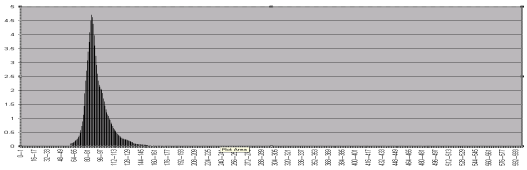

Figure 2.1.1c *Pitch distribution of an American English male speaker*

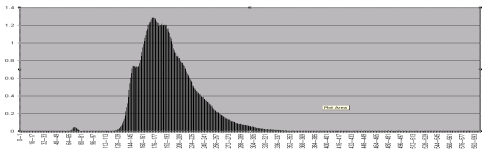

Figure 2.1.1d *Pitch distribution of an American English female speaker*

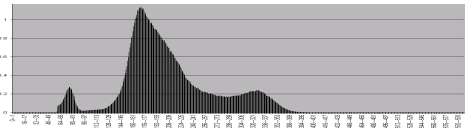

Figure 2.2.1e *Pitch distribution of a Cantonese female speaker*

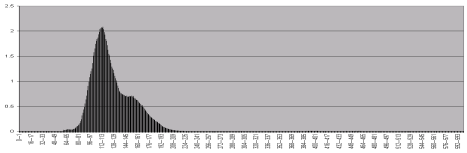

Figure 2.1.1f *Pitch distribution of a Cantonese male speaker*

The pitch distributions show that (1) there are small peaks below 100 Hz. It is the harmonic, or inherent fundamental frequency that is related to the masses of vocal folders; (2) the main pitch peaks are about twice of inherent frequency; (3) the both sides of main peaks are dissymmetry, particularly for Mandarin Chinese speakers, secondly for Cantonese speakers. There is only slight dissymmetry for American English speakers; (4) pitch ranges of female speakers are much boarder than that of male speakers.

Obviously, the pitch distributions do not follow Normal distribution and similar with gamma distribution. Figure 2.1.2 shows the Mandarin female speaker's pitch distribution and corresponding gamma analysis. In Bark scale, similar shapes of contour are happened.

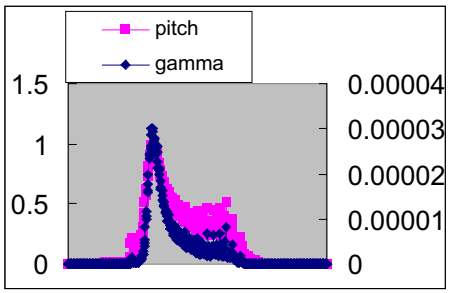

Figure 2.1.2 *Pitch distribution of the Mandarin female speaker and Gamma description*

2.2 Characteristics of pitch distribution in tone language

The discussion of this study focuses on syllable because in all Chinese dialects syllable is the smallest unit to carry lexical tones and in English syllable is the smallest unit to carry lexical stress. We use the speech data of Mandarin female speaker and American English speaker to discuss the differences of pitch distributions. Both English speaker and Mandarin speaker are broadcasters. The speech waveform was segmented into phone and syllable automatically. linguistics background students manually labeled break index and stress tier. As mentioned above, there is three break degrees: prosodic word, prosodic phrase and intonation phrase. The pitch distribution of English speaker is in a limited frequency range, while Mandarin and Cantonese speakers need boarder pitch range to realize lexical tones.

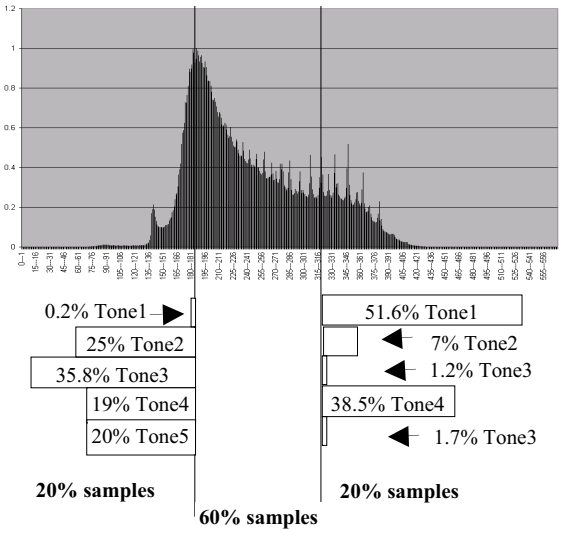

Figure 2.2 *Tone distribution in and pitch distribution of Mandarin female speaker's data*

There are four tones in Mandarin Chinese: high (tone1), rising (tone2), low (tone3) and falling (tone4). To simplify the problem, neutral tone is ignored. With the segmental annotation files of speech signal, we can determine tone attribute of each pitch sample. Figure 2.2 shows the relationship between tone distributions and pitch distribution. Tone distributions of 20% left samples and 20% right samples are plotted in this figure. Of course, the tone distribution is closely related to prosodic position. For instance, most of 0.2% tone1 in the left or lower frequency belongs to pre-boundary syllables.

2.3 Pitch analysis on Chinese speech data

Based on four tones of Mandarin Chinese, there are 15 tone pairs (tone3-3 -> tine2-3 because of tone sandih). There are two kinds of di-syllable pitch contexts, as Yi Xu defined, "compatible" and "conflicting". "Compatible" context is an environment in which adjacent phonetic units have identical or similar. "Conflicting" context is an environment in which adjacent phonetic units have very different value along phonetic dimension [3].

In the conflicting boundary, the F0 movement needs to change very quickly from one state to another. This time, pitch delay is always happen. In fact, only when the later target of preceding syllable has enough momentum of rising and the second syllable has low target, there exists real conflicting, such as 2-2, 2-3, and 3-4. In di-syllable words with voiced-unvoiced connection, the consonant initial of following interrupts pitch contour. We will focus on the cases that the initials of second syllable are voice. The voiced initial includes sonorant consonants m, n, l, r and vowel-initials or zero-initial, such as i, a, e, u and v. In this case, there exists entirely pitch

peak delay in second syllable. To clarify the problem, we only concentrate our discussion on tone2-3 di-syllable, which is composed of tone2 of preceding syllable and tone3 of following syllable. We call it tone2-3 pair here after. When the second syllable is extracted from the context, it will not be listened as low tone perceptively due to extra peak nearby onset. Among each pair groups, we class di-syllable words into two sub-groups, prominence/un-prominence as plotted in figure 2.3-1.

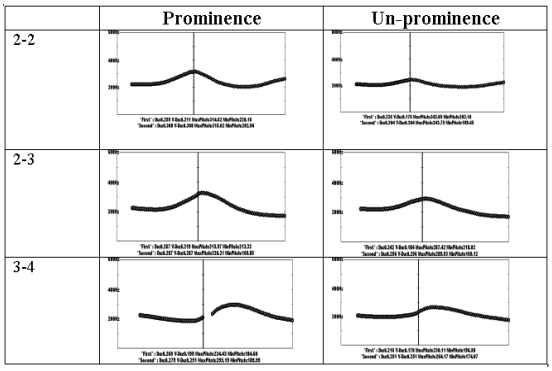

Figure 2.3-1 *Comparison of stressed and un-stressed tone pairs in conflicting case*

In Cantonese speech, there are nine tones. The Cantonese tone system can be drawn as figure 2.3-2 [4] based on five degrees. Tone1 - tone6 have long duration and tone7-tone9 have short duration. Although there are nine tones in Cantonese, the tone system is simpler than Mandarin. There are only rising tones. Different from Mandarin, there are three stop consonants affix in Cantonese syllable, which will interrupt pitch contour of di-syllable word. So the conflicting case in Cantonese is less crucial than in Mandarin speech.

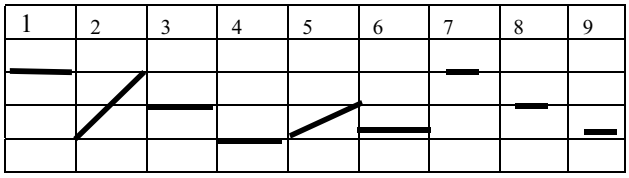

Figure 2.3-2 *Cantonese tone structure*

3. Tonal Alignment in Mandarin Chinese and English speech data

3.1. Evidence from Mandarin Chinese

Tonal alignment refers to the synchronization of pitch movement and segmental sequence. Yi Xu proposed that there exists underlying tone target in Mandarin and that pitch contours are probably implemented as integral dynamic targets rather than as sequences of static targets [5]. Xu's study [2] also shows that the variability in pitch contour can be attributed to the interaction among the underlying pitch targets, tone contexts and articulatory constraints, which represented by peak delay [6].

Xu conducted an experiment of maximum speed of pitch changes and represented the average minimum time needed to complete a pitch rising or falling as following equation (3.1-1 for rising and 3.1-2 for falling):

$$t = 89.6 + 8.7d \qquad (3.1\text{-}1)$$

$$t = 100.4 + 5.8d \qquad (3.1\text{-}2)$$

Where $F0_{\max}$ is maximum value at peak, $F0_{ref}$ is reference f0 value. d is semi-tone (St).

Here $d = 12\log_2 F0_{\max} / F0_{ref}$

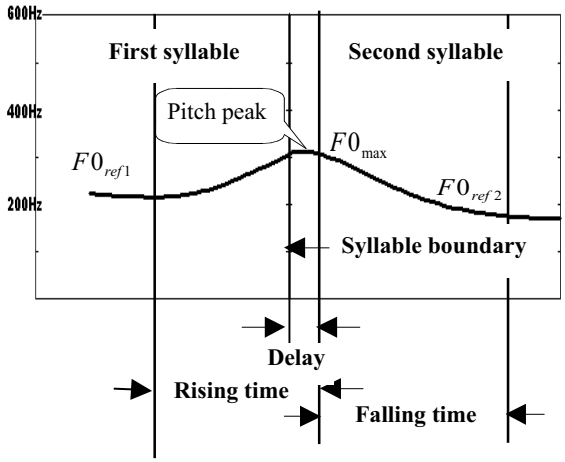

Figure 3.1 *Average pitch contours of tone2-3 pairs*

Table 3.1 *Predicted and tested durations for rising and falling*

	ΔF (St)	Predicted value (ms)	Real value (ms)
Rising	6.05	142	138
Falling	8.86	152	184

The average pitch contour of tone2-3 pair words in Mandarin speech corpus is shown in figure 3.1. According to Xu's equation, we calculated the predicted rising time and falling time. At same time we measured real values on the pitch contour. The results are shown in Table 3.1. In Figure 3.1, $F0_{ref1}$ is the pitch frequency value of first syllable at lowest point; 220Hz. $F0_{\max}$ is the pitch value of peak located at the highest point of second syllable, which equals 312Hz. There is about 17 ms delay from the syllable boundary.

The Mandarin female speaker takes 138 ms to finish a rising, and 184 ms to finish a falling. There is no significant difference between predicted value and real value in rising time, while the real value of falling time is longer than predicted value. In Xu's experiment, speakers try best to keep up the changes with the imitating pitch. In natural case it is un-necessary to keep the pitch movement so rapidly, based on economic rules. The similarity rising time may be due to the different preceding context in difference material.

In Mandarin female speech corpus, average duration of first syllable in di-syllable word is 244 ms, and average duration of second syllable in di-syllable word is 279 ms. As shown in table 3.1, 138 ms of rising time implies that about 100 ms in the front part of syllable is not used to perform rising. It can be expressed as affected by the underlying target of former syllable. For the following syllable, stating from the delayed peak, it takes 184 ms to finish a falling. Compared with the length of 279 ms of entire duration of second syllable,

it occupies more than half syllable duration (0.65) to reach low target of tone3. It will introduce a delayed pitch peak in the onset of second syllable inevitably. This part is perceptive noticeable. The speakers in this study, complete "rising – low" pattern will exceed average syllable duration in the speech rate of Mandarin speech corpus. If the speaker speeds up the speech rate, she will encounter more difficulty to realize a movement from rising target to low target even though she uses her utmost speech of pitch changes. The second syllable can be identified as tone3 only at the original context. If this syllable is used in another context for concatenation, it will bring failure because the isolated second syllable looses its intrinsic feature – low. This kind of failure is called alignment failure.

3.3 The evidence from English data

Figure 3.3 shows the average pitch contour of two associated syllables within an English word with a falling boundary tone. The preceding syllables have a pitch accent and following syllables are unaccented with voiced initials, such as /l, m, n/ and vowels. There is a pitch accent in this case, which is represented as a rising-falling tone. The pitch peak of this pattern is slightly ahead of segmental boundary, or located at the preceding syllable. There is 11 Sd deduce in the following syllable. The reason why alignment failure in English doesn't bring more troubles than Mandarin cases is that the lexical tone pattern in English is simpler than tone Languages. Only high and low lexical tone patterns in English will not form significant conflicting in tonal alignment with segments and does not occupy boarder frequency range. For this reason, there is enough freedom for an English speaker in tone dimension to realize high-level prosody. For instance, there are variable boundary tones for breaking. While in Mandarin Chinese, falling tone may be the unique mode in pre-boundary syllable. In English, pitch accent is always a result from joint effects, which relate to different prosodic levels [7].

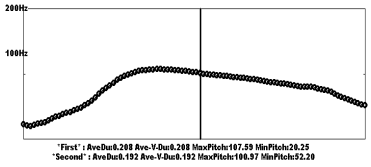

Figure 3.3 *Pitch accent curve extracted from the English speech corpus*

4. Conclusions and Discussion

4.1 Tonal alignment and concatenation cost of speech synthesis

The waveform concatenation speech synthesis [8,9] is still predominant in speech synthesis because it uses speech units excised from natural speech spoken by human being. Unfortunately, it is impossible to create a corpus, which can cover any text for any domains. So that discontinues will appear inevitably, even for a big corpus based TTS system. Therefore there are alignments among different articulators. The tonal alignment to segment is the hot topic in this area.

According to statistics on five years Chinese newspaper – People's Daily, the occurrence of voiced-voiced syllable adjacency is about 30% in 5,000 high frequently used words. The 5,000 high-frequency words will cover about 70% - 80% text. Same statistics is also conducted in English texts. It obviously cannot be ignored in concatenative TTS system.

4.2 Alignment failure in tone language

The speech data in this study comes from a natural TTS speech corpus. The results about alignment failure demonstrate the consistency with Xu Yi's results, which are based on designed test sentences. Under specific architecture of speech synthesizer, the quality of speech output depends on acoustic-phonetic details.

In English speech, the alignment failure will not bring too much trouble because there are only high and low features to realize pitch accent in English. Native English speakers will use variable boundary tones to break an utterance. It is fair for Mandarin Chinese speakers to use a simple falling tone in pre-break because they have to apply complex lexical tones in same dimension. The investigations on mechanism of tonal alignment will help us to deal with concatenation details.

The alignment failure will happen in complex tone systems, or tone language. The significant dissymmetry in mandarin and Cantonese concludes that the boarder pitch range in frequency domain dose not to meet with the needs of non-linear frequency changes in perception level, but due to the realization of complex tone systems. The alignment failure always happens in rising tone related cases because there is a strong momentum at this time. We are still wondering if the pitch movement in higher frequency plays the similar roles in lower frequency range.

5. References

[1] Yi Xu, Q. Emily Wang, 2001. *Pitch targets and their realization: Evidence from Mandarin Chinese,* Speech Comunication 33: 319-337.

[2] Yi Xu & Xuejun Sun, 2002. *Maximum speed of pitch change and how it may relate to speech*, J. Acoust. Soc. Am. 111(3), March.

[3] Yi Xu, 1995. *Production and perception of coarticulation tones*, J.Acoust. Soc. Am. 95(4).

[4] Tien-Ying Fung and Hellen M. Meng, 2002. *"The effect of tonal context on Cantonese concatenative speech synthesis",* ISCSLP'2002. pp.22-24.

[5] Yi Xu, 1998. *Consistency of tone-alignment across different syllable structures and speaking rate*, Phonetica, pp55:179-203.

[6] Yi Xu, 2001. *Fundamental frequency peak delay in Mandarin,* Phonetica, 58:26-52.

[7] Jan P.H. van Santen, 2002. *Quantitative Modeling of Pitch Accent Alignment*, Proceedings of the 1st International Conference on Speech Prosody, pp. 107-114.

[8] A. Hunt and A. Black, 1996. *Unit selection in a concatenative speech synthesis system using a large speech database*, in Proc. ICASSP '96, Atlanta, GA: 373-376.

[9] Alan W Black, and Nick Campbell, 1995. *Optimising Selection of Unit from Speech Database for Concatenative Synthesis*, Eurospeech.