THREE LEVELS OF TUNE-TEXT RELATIONS

YI XU
The University of Chicago, USA
E-mail: xuyi@uchicago.edu

ABSTRACT

This paper proposes that to understand various issues concerning tune-text alignment, three levels of timing relations need to be recognized, namely, *underlying association*, *target synchronization*, and *surface alignment*. While the link between underlying association and target synchronization may be stipulated by various phonological rules, the link between surface alignment and target synchronization can be better understood only through the recognition of various articulatory constraints. Most importantly for experimental research, surface alignment only *indirectly* reflects the alignment at the other two levels, and thus should not be taken as intended by the speaker as such.

1. INTRODUCTION

As speech analysis technology advances, we can observe more clearly than ever acoustic details in the speech signal. This has allowed us, among other things, to examine the exact time alignment of different acoustic events in an utterance [1, 2, 3, 5, 9, 18, 19, 20] (to cite just a few). The directly observable alignment patterns, however, do not necessarily tell us whether a particular alignment pattern is intended by the speaker as such, or is the consequence of the interaction between different alignment-affecting factors. In this paper, I would like to propose that to fully understand the issues relating to tune-text alignment, there is a need to recognize three levels of timing relations, namely, *underlying association*, *target synchronization*, and *surface alignment*. I would like to argue that, though intimately related, these three levels have different properties, and are linked by different mechanisms.

2. UNDERLYING ASSOCIATION

Underlying association refers to how various linguistically functional components are associated with each other in time sequence. These components include phonological units such as consonants, vowels, lexical tones, pitch accents, etc. Consonants and vowels are presumably combined into syllables at this level, and tones and pitch accents seem to be associated with the syllable [9, 18, 19, 22]. In a tone language such as Mandarin, for example, the smallest meaningful unit, known as the morpheme, is the size of a syllable, which consists of both segments and a tone. The association of tone and syllable is not always straightforward, however. In some tone languages, a syllable may be deleted by a certain phonological process, but the tone associated with it is still functional and is thus required to remain [15]. Such a tone then becomes "floating" and still needs to be associated with a syllable in one way or another [10]. In some other languages, certain tones may seem to have changed their identity in a particular context. In Mandarin, for example, the L (Low) tone seems to change into a form that is perceptually indistinguishable from the R (Rising) tone when it is followed by another L tone [16]. In this case, either the tonal association at this level is changed, or a phonological process has generated a pitch target that is indistinct from that of the R tone. Underlying association of tonal components, however, cannot be observed directly. In fact, the only timing relation that is directly observable is the *surface alignment*, as will be discussed next.

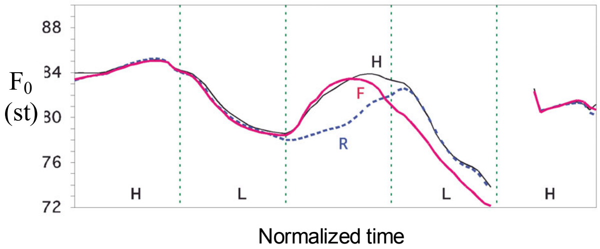

Fig. 1. Mandarin F (Falling), H (High) and R (Rising) tones between two L (Low) tones. Vertical lines indicate onset of syllable initial consonants. Adapted from [19].

3. SURFACE ALIGNMENT

Surface alignment refers to how directly observable F_0 events, such as peaks, valleys, and contours, are positioned in time relative to segmental landmarks, such as consonant closure onset and release, vowel onset and offset, and so on. It may at first sight seem reasonable to assume that observable pitch alignments must be intended by the speaker, at least subconsciously. Take the lexical tones in Mandarin as an example. In Fig. 1 we can see that, when surrounded by two L tones, the H, F and R tones seem to differ from one another in the alignment of F_0 peaks relative to syllable boundaries: The F tone has the earliest peak within the syllable, the H tone has a later peak within the syllable, whereas the R tone has a peak after the end of the syllable. Such alignment differences may appear quite apparent when displayed in this way. But are they the essence of the tonal distinction in Mandarin [5]? Another way of observing the data may suggest a different interpretation. Fig. 2 shows the same three tones in the same sentence position, but this time each with four different preceding tones displayed in the same plot. What we can see now is how the same tone is produced after different preceding tones. While the four tones in syllable 2 end with very different offset F_0, the contours in syllable 3 in each panel all gradually converge, not to a single peak or valley,

ISBN 1-876346-48-5 © 2003 UAB

but to a certain linear configuration, as indicated by the straight dashed lines: a fall in the top panel, a high-level in the middle panel, and a rise in the bottom panel. Seen in this way, the F_0 peak alignment patterns, which appear to separate the three tones so nicely in Fig. 1, no longer seem to be the most important characteristics of these tones. In fact, in the case of the F tone (top panel of Fig. 2), the location of the peak varies with the preceding tone: the lower the ending F_0 of the preceding tone, the later the peak. The same is also true in the bottom panel in regard to the alignment of the F_0 valley in the R tone.

In addition to the location of the converging F_0 configuration, there is another alignment that also seems to be quite constant in Fig. 2. That is, F_0 movements toward the converging configuration in each tone consistently start from the onset of syllable 3. In the top two panels, the movements toward the linear fall or the level high all start as a rise from where the previous tone ends. In the bottom panel, the movements toward the linear rise start from the syllable onset either as a rise or as a fall depending on what the offset F_0 of the previous tone is. Putting the two kinds of alignment patterns together, it becomes apparent that the F_0 contour of an entire syllable consists of a movement toward some ideal F_0 configuration. In other words, if the production of a tone is an implementation of a target configuration, such implementation seems to be fully synchronized with the syllable with which the tone is associated. This observation leads us to the middle level of tune-text relations: *target synchronization*.

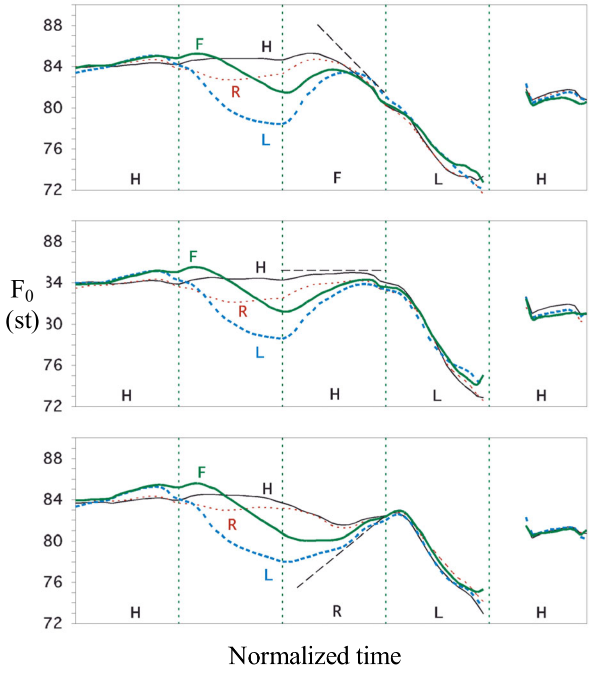

Fig. 2. Mandarin F, H and R tones (in syllable 3) followed by the L tone but proceeded by four different tones. Note how each tone seems to converge to a linear configuration indicated by the dashed straight line. Adapted from [19].

4. TARGET SYNCHRONIZATION

Target synchronization refers to how different phonetic targets are implemented coordinately in production. *Phonetic targets* are defined as the smallest articulatorily operable units associated with various phonological elements. In the case of F_0 production, the smallest articulatorily operable unit is presumably a pitch target, which is a linear configuration that is either static or dynamic [22]. The recognition of this level is necessary for several reasons. First, underlying phonological elements are often not articulatorily operable. For instance, the L tone in Mandarin, as a phonological unit, is not readily producible because speakers would not know whether to say it with a falling-rising, low level, or rising configuration, unless they also know the context in which it is to be said. In the left graph in Fig. 3, syllable 2 is associated with the L tone while syllable 1 is associated with four different tones. The L tone realized in syllable 1 is apparently not very different in shape from the R tone in the same syllable, although the two differ somewhat in overall height. Furthermore, the L tone in the right graph of Fig. 3, which is produced in isolation, has a final rise that is totally absent in the L tone in syllable 2 in the left graph, which is not produced in isolation. Thus the L tone in Mandarin seems to have at least three distinct target configurations, low+rise in isolation, rise before another L tone, and low in other cases. At the same time, however, we cannot ignore the obvious fact that these targets are all linked to the L tone in the lexicon, and the link seems to be governed by phonological rules that are somewhat stipulative. Thus there must exist an intermediate level, at which are phonetic targets that are linked to more abstract units at the underlying phonological level on the one hand, and to the acoustic events at the surface level on the other.

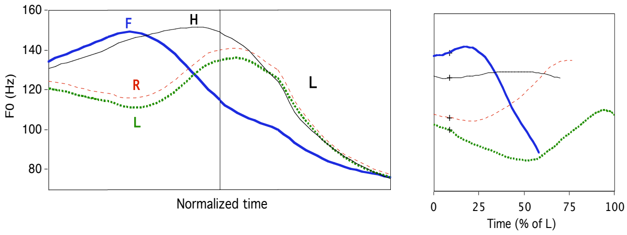

Fig. 3. Left: Mandarin L tone after four different tones, produced in a carrier phrase. Right: four Mandarin tones produced in isolation. Adapted from [17].

The second reason for recognizing an intermediate level is that, as discussed earlier, surface alignment patterns are often unlikely to have been intended by speakers as such. It is not likely, for example, that the F_0 rises in syllable 2 of the middle panel of Fig. 2 are intended as rising configurations. Rather, they are more likely to be consequences of speakers' effort to realize a high-level configuration when the F_0 movements have to start from levels lower than the height of the targeted configuration. Thus the targeted configurations often seem to be "hidden" behind what can be observed on the surface. Without recognizing their existence, therefore, we would have missed the consistency across the seemingly inconsistent surface patterns. Furthermore, this consistency cannot be fully understood until the sources of the variance are identified. One of the sources seems to be that changing vocal pitch takes time. This is because pitch is produced by the larynx, which, being a physical device, cannot change its state instantaneously. Of course, it could be possible that the time it takes to change (raise or lower) pitch is so short that it can be safely ignored in most cases, as have been argued before [2, 14]. But we would not be sure about this until the maximum speed of pitch change has been accurately assessed.

The third reason for recognizing the level of target synchronization is probably the least obvious, but no less important. It has to do with the fact that the implementation of the phonetic targets must be coordinated, and there is a limit as to how freely different articulatory movements can be related to each other temporally. Its importance becomes clearer once we are more certain about the maximum speed of pitch change, which will be discussed next.

5. MAXIMUM SPEED OF PITCH CHANGE

Maximum speed of pitch change refers to the highest speed at which speakers can voluntarily raise or lower pitch. This physiological limit has been examined in several studies [12, 13, 21]. The latest study investigated this limit for both Mandarin and English speakers (male and female), and obtained data in a format that can be easily compared to those from real speech [21]. A common finding of all these studies is that the maximum speed of pitch change is closely related to the magnitude of the pitch change. For example, the following relations between the maximum speed of pitch raising and lowering were observed in [21]:

$$s = 10.8 + 5.6\,d \qquad \text{(raising)} \qquad (1)$$

$$s = 8.9 + 6.2\,d \qquad \text{(lowering)} \qquad (2)$$

$$t = 89.6 + 8.7\,d \qquad \text{(raising)} \qquad (3)$$

$$t = 100.4 + 5.8\,d \qquad \text{(lowering)} \qquad (4)$$

where s is the average maximum speed of pitch change in semitones per second (st/s), t is the amount of time (ms) it takes to complete the pitch shift, and d is the size of pitch shift in st. Equations (1)-(4) suggest that, first, the constraint of maximum speed of pitch change inevitably would leave an imprint on the surface F_0 patterns. For example, taking another look at the middle panel in Fig. 2, we note that after the L tone in syllable 2, F_0 has to increase 5-6 semitones to reach the height of the H tone. According to equation (3), an average speaker needs at least 133-142 ms to complete this pitch elevation. Since the average duration of syllable 3 is about 181 ms [19], much of the syllable duration has to be used to accomplish this elevation even if the speaker uses the fastest speed. Furthermore, equations (1)-(4) also suggest that in many occasions, the maximum speed of pitch change is indeed approached in Mandarin. The top panel of Fig. 2 shows that to realize the targeted fall for the F tone, F_0 following the L tone needs to first go up and then drop down. This maneuver means that two consecutive F_0 movements are made within syllable 3: rise and then fall. The size of the rise is about 5 semitones and the fall 2.5 semitones. From equations (3) and (4), the combined time needed for completing these two movements is about 248 ms, which is longer than the average duration of syllable 3 (181 ms)! Of course, speakers could not have exceeded the maximum of speed of pitch change, and they did not, based on the calculation of [21]. Nevertheless, it is precisely when producing the dynamic tones that Mandarin speakers seem to have approached the maximum speed of pitch change [21]. [21] also notes that the fastest pitch movements in Dutch as reported in [2] also have probably approached the maximum speed of pitch change.

6. COORDINATION OF BODILY MOVEMENTS

As found in studies on limb movements, to carry out concurrent bodily movements, performers have very few choices in terms of timing relations between involved movements [6]. At relatively slow speed, the phase angle between two movements has to be either 180º, i.e., starting one movement as the other is half way through its cycle, or 0º, i.e., starting and ending the two simultaneously. At high speed, however, only the 0º phase angle is possible. The findings of [6] thus indicate that there is a strong tendency against any phrase angles other than 180º and 0º. For speech, there has been evidence that the syllable functions as a coordinative structure to which many articulatory movements are aligned [4, 8]. In the case of lexical tones the tonal targets thus have to be aligned to the syllable either at 180º or 0º phrase angle. If tones are directly associated with the syllable in the lexicon as in Mandarin, there would hardly be any conceivable motivation for maintaining a constant 180º phase angle between syllables and the tonal targets. This leaves 0º phase angle the only viable choice for the speaker. Indeed, existing data strongly suggest that a full synchrony of the two is enforced [18, 19, 20]. As we have seen earlier, the implementation of dynamic tones such as R and F in Mandarin often requires the maximum speed of pitch change to be approached, which would imply much articulatory effort on the part of the speaker. The principle of economy of effort [11] may suggest that speakers might want to start the implementation of the pitch target earlier or end it later to avoid spending excessive effort. However, as can be seen in Fig. 4, they instead consistently start the movement at the syllable onset and end it at the syllable offset. As a result, it is the peak F_0 that appears to be compromised, being much lower in the L F than in the H F sequence. In other words, the synchrony constraint seems to have prevailed over both the need to avoid excessive effort and the need to maximally realize a linguistic category.

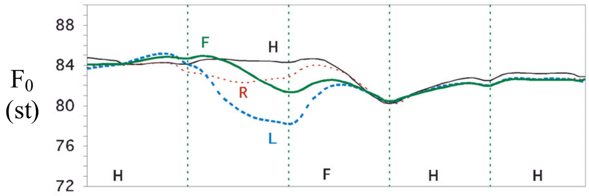

Fig. 4. Mandarin F tone followed by the H tone but proceeded by four different tones. Adapted from [19].

The bottom panel of Fig. 2 shows two more phenomena that have to do with synchronization. The first is that the F_0 peak that is apparently associated with the R tone occurs in the following syllable which carries the L tone. This "peak delay" is presumably the consequence of F_0 rising sharply at the end of syllable 3 [20, 22]. Due to inertia, the laryngeal movement cannot immediately reverse the rise albeit the implementation of the L tone in syllable 4 is likely to have begun at the syllable onset. The second phenomenon is that when syllable 2 carries the L tone, F_0 in syllable 3 rises at first only slowly, but then increases its speed in the later portion of the syllable. Thus there seems to be a partial delay of the sharp rise. A similar phenomenon was reported in [18] in which F_0 alignment at different speaking rates was examined. It was found that when the syllable duration increased with decreased speaking rate, the rise in the R tone moved later relative to syllable onset. Both the delayed peak and the late rise seem to indicate that the pitch target executed in the R tone is a simple rise rather than a sequence of low and high, as assumed by many phonological analyses. The late rise in the R tone is reminiscent of the finding of another study on human interlimb movements [7]. It is known that it takes longer for a hand or finger to reach a target that is more difficult, i.e., farther away or smaller in size. When performing a bimanual task in which the hands or fingers need to reach separate targets of unequal difficulty, subjects nevertheless reach the two targets simultaneously. They do so by slowing down the hand reaching the easier target [7]. This seems rather like what speakers do with the R tone at slow speaking rate: when the syllable is too long, they delay the onset of the rise so that a sharp slope is still achieved by the end of the syllable. Thus pitch targets and syllables appear to be synchronized both when time pressure is too high and when it is too low, although apparently different strategies are employed to guarantee the synchrony.

7. CONCLUSION

To summarize, I have demonstrated that surface alignment patterns cannot be directly linked to underlying phonological association. Rather, there must exist an intermediate level at which are phonetic targets that are linked to more abstract units such as lexical tones by phonological rules. Phonetic targets differ from the more abstract phonological units in that they have to be readily executable articulatorily. During production, the targets are implemented by an articulatory system that has various inherent constraints. These constraints, including the maximum speed of pitch change and coordination of articulatory movements, inevitably leave various imprints on the alignment patterns directly observable in the acoustic signal. To understand any surface alignment patterns, therefore, it is imperative to recognize the articulatory mechanisms that are responsible for implementing phonetic targets.

ACKNOWLEGEMENT

This work is supported in part by NIH Grant DC03902.

REFERENCES

[1] Bruce, G., "Developing the Swedish intonation model", *Lund University, Dept. of Linguistics Working Papers,* **22**, pp.51-116, 1982.

[2] Caspers, J. and van Heuven, V. J., "Effects of time pressure on the phonetic realization of the Dutch accent-lending pitch rise and fall", *Phonetica,* **50**, pp.161-171, 1993.

[3] D'Imperio, M., "Language-Specific and Universal Constraints on Tonal Alignment: The Nature of Targets and "Anchors"", *Proc. 1st Int. Conf. on Speech Prosody*, Aix-en-Provence, France, pp.101-106, 2002.

[4] Fujimura, O., "The C/D model and prosodic control of articulatory behavior", *Phonetica* **57**, pp.128-138, 2000.

[5] Gårding, E., "Speech act and tonal pattern in Standard Chinese", *Phonetica,* **44**, pp.13-29, 1987.

[6] Kelso, J. A. S., Holt, K. G., Rubin, P. and Kugler, P. N., "Patterns of human interlimb coordination emerge from the properties of nonlinear limit cycle oscillatory process: theory and data", *J. Mot. Behav.,* **13**, pp.226-261, 1981.

[7] Kelso, J. A. S., Southard, D. L. and Goodman, D., "On the nature of human interlimb coordination", *Science,* **203**, pp.1029-1031, 1979.

[8] Krakow, R. A., "Physiological organization of syllables: a review", *J. Phonetics* **27**, pp.23-54, 1999.

[9] Ladd, D. R., Mennen, I. and Schepman, A., "Phonological conditioning of peak alignment in rising pitch accents in Dutch", *J. Acoust. Soc. of Am.,* **107**, pp.2685-2696, 2000.

[10] Laniran, Y., "Implementing a floating tone", *Proc. 71st Ann. M. of Ling. Soc. of Am.*, Chicago, pp.59-60, 1997.

[11] Lindblom, B., "Explaining phonetic variation: A sketch of the H&H theory", *Speech Production and Speech Modeling*. W. J. Hardcastle and A. Marchal. Kluwer, Dordrecht, The Netherlands, pp.413-415, 1990.

[12] Ohala, J. J. and Ewan, W. G., "Speed of pitch change", *J. Acoust. Soc. of Am.,* **53**, pp.345(A), 1973.

[13] Sundberg, J., "Maximum speed of pitch changes in singers and untrained subjects", *J. Phonetics,* **7**, pp.71-79, 1979.

[14] 't Hart, J., Collier, R. and Cohen, A., "*A perceptual Study of Intonation — An experimental-phonetic approach to speech melody"*, Cambridge University Press, Cambridge, 1990.

[15] Tadadjeu, M., "Floating tones, shifting rules, and downstep in Dschang-Bamileke", *Studies in African Linguistics,* **Supplement 5**, pp.283-291, 1974.

[16] Wang, W. S.-Y. and Li, K.-P., "Tone 3 in Pekinese", *J. Speech Hear Res.,* **10**, pp.629-636, 1967.

[17] Xu, Y., "Contextual tonal variations in Mandarin", *J. Phonetics,* **25**, pp.61-83, 1997.

[18] Xu, Y., "Consistency of tone-syllable alignment across different syllable structures and speaking rates", *Phonetica,* **55**, pp.179-203, 1998.

[19] Xu, Y., "Effects of tone and focus on the formation and alignment of F_0 contours", *J. Phonetics,* **27**, pp.55-105, 1999.

[20] Xu, Y., "Fundamental frequency peak delay in Mandarin", *Phonetica,* **58**, pp.26-52, 2001.

[21] Xu, Y. and Sun, X., "Maximum speed of pitch change and how it may relate to speech", *J. Acoust. Soc. of Am.,* **111**, pp.1399-1413, 2002.

[22] Xu, Y. and Wang, Q. E., "Pitch targets and their realization: Evidence from Mandarin Chinese", *Speech Commun.,* **33**, pp.319-337, 2001.