THE MISSING LINK BETWEEN ARTICULATORY GESTURES AND SENTENCE PLANNING

Chilin Shih**, **Greg Kochanski and **Su-Youn Yoon***

University of Illinois at Urbana-Champaign and Oxford University*
cls@uiuc.edu, greg.kochanski@phon.ox.ac.uk, syoon9@uiuc.edu

ABSTRACT

The motivation of this paper is to build a bridge between phonology and phonetics with implementation models. The challenge is to explain a wide range of phonetic forms in diverse speaking styles, including laboratory speech, spontaneous speech, fluent and non-fluent speech and model them as orderly variations of one coherent communication system. In the paper, we will focus on the effect of sentence planning on articulatory gestures.

Data from spontaneous speech provides strong evidence for anticipatory effects and partial reduction effects. How and when they occur reflect the speaker's sentence planning strategies. It is hypothesized that much of the discrepancy between laboratory speech and spontaneous speech can be accounted for with a model that can represent these effects, such as the weights used in the Stem-ML model.

Keywords: anticipatory effect, pre-planning, prosodic variations, partial reduction, weighted models

1. INTRODUCTION

The motivation of this paper is to build a bridge between phonology and phonetics with implementation models [11]. The challenge is to explain a wide range of phonetic forms in diverse speaking styles, including laboratory speech, spontaneous speech, fluent and non-fluent speech and model them as orderly variations of one coherent communication system.

In the paper, we will focus on the effect of sentence planning on articulatory gestures and address two seemingly un-compromisable positions: On the one hand, there seems to be a precise mapping relationship between phonological representations and acoustic events, supported mostly by experimental works and laboratory speech where speech sounds come in pre-defined sequence and there are precise alignment patterns between prosodic functions and forms as well as between segmental and prosodic events (See [14].) On the other hand, experience from speech recognition and classification tasks suggests a much more complicated situation. Phonologically specified targets and gestures may be exaggerated, reduced, fused, or deleted in spontaneous speech, leading to considerable amount of asynchrony [8]. It is not only difficult to establish unambiguous acoustic landmarks and alignment patterns that would help the identification of speech segments, tones, accents and other prosodic events, very often the assumed anchors of such events disappear all together.

A missing link between the two worlds is a formal mechanism that models speech variation and treats laboratory speech and spontaneous speech as one coherent communication system. The *Soft Template Markup Language (Stem-ML)* model [3, 4] is a small step toward the understanding of this problem. The model uses weight on every target/gesture to simulate the effect of exaggeration and reduction, especially partial reduction, which cannot be handled by target deletion and requires a way to resolve conflicts. The weights represent the speaker's balancing act to meet two conflicting demands in speech communication: for ease of production, the speaker would like to minimize articulatory effort; for ease of perception, the speaker would need to maintain production accuracy. On this point, Stem-ML is conceptually similar to the H & H model [7]. The two models differ in how weights are used and modeled. The H & H model applies weights to databases to capture differences in speaking styles, while Stem-ML uses weights to capture production variations from one word/syllable to the next. The interpretation of the weights is modeled in Stem-ML with knowledge of the global trajectory defined by articulatory targets. On this point, Stem-ML differs from the *Parallel Encoding and Target Approximation (PENTA)* model [15, 14] which makes strong theoretical claims disallowing a look-ahead component. Production variations in PENTA are explained with a model of target approximation with knowledge of the previous target and the current target.

In this context, this paper examines one of the crucial differences between Stem-ML and PENTA: whether there is a place for anticipatory effects in a speech production model. This concept is directly related to the way surface variations are modeled. Data from spontaneous speech provides strong evidence for anticipatory effects, supporting speech production models with preplanning and look-ahead. A study of fluency in spontaneous speech [16] also supports such a model: fluent speech is characterized by dynamic duration control that reflects an effective communicative strategy.

It is hypothesized that the discrepancy between laboratory speech and spontaneous speech can be accounted for with a model that can represent different sentence planning strategies, such as the weights used in the Stem-ML model.

2. THE PENTA ASSUMPTION

The PENTA model disallows look-ahead:

> "No anticipatory execution–The movement toward a target does not start until the movement toward the preceding one is over" [14].

This is a strong prediction, denying the possibility of anticipatory effects in speech. In that model, the discrepancy between intended tonal targets and their surface realization is explained by the sluggishness of articulatory movement as it approaches the current target from the previous one. Furthermore, the PENTA model treats the syllable as a synchronization unit. Its prediction is that speech production unfolds one syllable at a time. Each syllable has its own specified target; the articulatory movement starts at the syllable onset, and approaches the specified target at the end of the syllable.

Two explanations for some of the anticipatory effects reported in the literature are provided by the PENTA model [14]. First, it is well established that an initial consonant may be colored by the following vowel [9, 1]. For example, the initial consonant [s] is unrounded in the word *see* but rounded in the word *Sue*. The rounding of the consonant seems to anticipate the rounding of the vowel. This phenomenon is not treated as an anticipatory effect in PENTA. Rather, it is explained as co-production of consonant and vowel by having the vowel gesture starting at the syllable onset. The rounding of the consonant is then seen as the result of overlapping vowel and consonant gestures. While this is a plausible explanation for short-range anticipatory coarticulation, it is not clear how it can account for the apparent anticipation changes two or more syllables in the future [13].

Second, changes in formant or f_0 trajectories are often made inside the preceding syllable. This phenomenon is explained in PENTA by redefining the syllable boundary. The beginning of the syllable is defined to be the point when an articulator starts moving toward the next target, which comes earlier than the boundary defined by acoustic properties. This treatment can only explain anticipation effects that are confined to a fraction of a syllable, and once assumed, cannot easily be turned off. The model, then, becomes less capable of explaining delays of tonal turning points.

In contrast, pre-planning is an important aspect of the STEM-ML model. Its characteristic prediction is that local reductions and even inversions of articulations should sometimes be observed, when the reduction/inversion is part of a global optimization. In the following, research findings will be reviewed to suggest that anticipatory effects are common in speech. Data from spontaneous speech provides strong evidence for anticipatory effects and pre-planning that goes beyond consonant-vowel co-production and the uncertainty of syllable boundary placement. Furthermore, it may be one of the best features that characterizes fluent spontaneous speech. A formal model proposed in Stem-ML that uses weights to control speech production is an effective way to account for these observed variations.

3. ANTICIPATORY EFFECTS IN SPEECH

Mandarin Chinese is a tone language that uses pitch shape and pitch height to distinguish lexical meaning. For example, the word pair *mai3* "to buy", and *mai4* "to sell" are differentiated by tones, the word "to buy" has a low-falling-rising tone (tone 3), and the word "to sell" has a high falling tone (tone 4). In theory, most syllables in a sentence, if not all, come with lexical tones that would specify the f_0 trajectory. Given the relatively unambiguous state of lexical tones, the expected sequence of tonal targets for a sentence should be predictable from the text. However, there often are considerable discrepancies between the expected target and the surface realization [4, 5]. Figures 1 and 2 show two such examples. From 10 to 15% of syllables in conversational speech show distorted tone production like this [4]. The F_0 tracks are plotted as a function of time. There is no phrasal boundary in either case. Figure 1 shows a lexical falling tone being realized with a rising shape in the phrase *mai3 mai4 jiu4-che1 wo[2] gan3* "Buy and sell used cars, I dare (to do it)." The second syllable *mai4* "to sell" with a lexical falling tone is produced with a rising trajectory. Figure 2 shows a lexical rising tone being realized with a falling shape in the phrase *zai4 ming2-nian2 nian2-di3* "at the end of next year." The third syllable *nian2* "year" with a lexical rising tone is produced with a falling trajectory, in contrast with the same morpheme with the expected tone shape occurring immediately after. These examples can be explained as local tonal distortion with global optimization in favor of a smooth movement trajectory. The planning of such a trajectory necessarily involves anticipating the tonal targets of upcoming syllable(s).

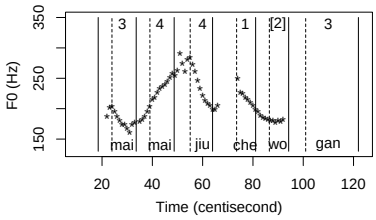

Figure 1: The F_0 track of the phrase *mai3 mai4 jiu4-che1 wo[2] gan3* "Buy and sell used cars, I dare (to do it)," plotted as a function of time. The second syllable *mai4* "to sell" with a lexical falling tone is produced with a rising trajectory.

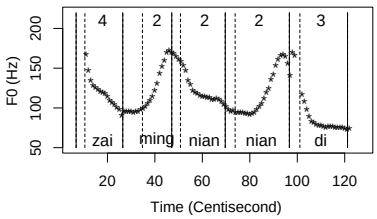

Figure 2: In this phrase, *zai4 ming2-nian2 nian2-di3* "at the end of next year," the third syllable *nian* with a lexical rising tone is produced with a falling trajectory in anticipation of the following tone, which starts low.

These two phrases represent a class of examples that are not compatible with the prediction of PENTA. For instance, in Figure 1, the boundary between syllables 1 and 2 would have to fall earlier on the basis of the valley in the f_0 contour. This might be accounted for by the PENTA re-definition of syllable boundaries, but then it is hard to explain the delayed boundary between syllables 2 and 3. Also, the presumed falling tonal target for the tone 4 on the second syllable is not approached by the end of the

syllable. To explain syllable 2 via PENTA would require the end point of the target to somehow be shifted 100 Hz upwards in frequency, with no obvious explanation.

This case may be explained by target deletion followed by interpolation in a ToBI like model [10], but an interpolation step also brings in information from future syllables, and is therefore anticipatory.

Further, complete deletion does not seem sufficient to account for all production variation. The distortion in the third syllable in Figure 2 is a case of partial reduction where the surface f_0 trajectory is heavily influenced by the global context but there are traces of the lexical tone gesture showing resistance to the falling trend created by surrounding tones. It is not strong enough to result in an intended rise and only manages to change the slope of the fall. This phenomenon is similar to the findings of [2] where the articulatory gesture of [t] is present in cases that would have been labeled deletion because the acoustic landmark of the [t] is not present due to the lack of consonantal closure. Partial reduction cannot be accounted for by deletion, under-specification or tonal interpolation and requires a model that can explain a range of production variations.

These examples show that several of the assumptions of PENTA are frequently violated: targets are not always approached closely, syllable boundaries do not always correspond to shifts from one tonal target to another, and the surface f_0 trajectory shows simultaneous effects from the preceding, current and following tones.

4. REDUCTION FOR GLOBAL OPTIMIZATION

The long-range effects are particularly interesting because most models of coarticulation can accommodate long-domain effects only if there are no intervening phones that compete for an articulatory target. The Stem-ML model allows target configuration that may only exist in one's mind, including, for example, targets that overlap with conflicting demands. Each target comes with a weight specifying how strongly it should be implemented. By using weights, the Stem-ML model offers a way to link phonology with phonetics [12], or invariant target with a wide range of surface variations [5].

Figure 3 illustrates conceptually how this model works. The mathematical models were provided in [3]. Starting with hypothetical targets such as three rising tones, represented by the dashed lines, different surface forms are generated depending on the weights on each target, which are given in the parenthesized numbers. In these three examples, the weights on the first and last syllables are kept constant and the middle one changes (1.0, 0, and 0.5). The realization of a target depends on the weight of the target and its neighbors. A target with a relatively strong weight is realized, probably with some small compromises with its neighbors, as seen in the middle syllable of (a). A target with zero weight is equivalent to deletion of the target, as shown in the middle syllable of (b). An intermediate weight results in partial reduction and extensive compromise with its neighbors, as shown in the middle syllable of (c). An intermediate weight gives a trajectory intermediate between (a) and (b), showing the influence of both lexical tone and the tonal trajectory

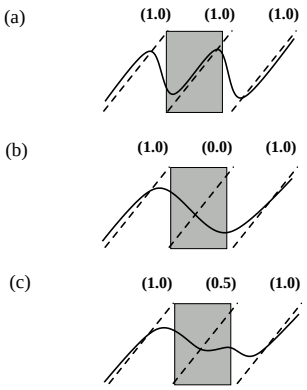

Figure 3: The realization of a target depends on its weight and the weights of its neighbors. A target with a relatively strong weight is realized in a form fairly close to its target, as seen in the middle syllable of (a). A target with zero weight is equivalent to target deletion, as shown in the middle syllable of (b). An intermediate weight (0.5) results in partial reduction, as shown in the middle syllable of (c).

of the surrounding tones. All three cases, as well as numerous intermediate stages, are found in natural speech. Example (c), for example, captures the pattern shown in Figure 2.

The tone examples in Figures 1, 2 as well as the simulation in Figure 3 illustrate a pattern where local reduction is predictable in context as the result of global optimization. Weights on targets are used to control the production of surface variations [5]. The interpretation of the weights implies pre-planning.

5. THE MEASUREMENT OF FLUENCY

The previous section shows how the concept of weights is used in the Stem-ML model to account for tonal variations in a sentence. This idea can be generalized to explain the difference between different styles of speech such as fluent vs. non-fluent speech or conversational vs. laboratory speech. In (a) of Figure 3, the weights assigned to different syllables are comparable, hence each tone is produced similarly. In (b) and (c) of Figure 3, the values of the weights span a wide range, which lead to different production of the same tone. In [4], we showed that the estimated weights trained with the Stem-ML model from news reading correlate with mutual information and duration, and further reflect the discourse structure, part of speech categories, word length, and the rhythmic structure.

In a recent study, Yoon [16] investigated the duration patterns of the word *shi4* "to be" in a spontaneous Mandarin speech corpus produced by native Mandarin speakers and Shanghai speakers [6]. Shanghai is a Wu dialect which is not mutually intelligible with Mandarin Chinese. These speakers learned Mandarin with different level of proficiency. The word *shi4* is chosen because of its high frequency in the corpus and its distribution in a wide variety of contexts. Yoon found a difference in the sentence planning process among native speakers of Mandarin, Shanghai speakers who are ranked as fluent and

as non-fluent. In Figure 4, the averaged duration values (in msec) is plotted as a function of the log conditional probability, which estimates how likely the word *shi4* will occur given the preceding context. Native speakers' data is plotted in solid lines, non-native fluent speaker in short dashed lines and non-native non-fluent speaker in long dashed lines. All speakers show some sensitivity to context and use shorter duration when the word is predictable and longer duration when it is not. However, the native speaker group is most effective in controlling speaking rate to reflect the information structure of the spoken message. If the word *shi4* is less predictable, the listeners would need more help. In this case, the duration values of the native group (178 msec) is not only longer than words in more predictable context, it is also longer than the production by non-native speakers. In contrast, the non-fluent group is the least effective in applying this strategy. Comparing to the other two groups, they use shorter duration (136 msec vs. 178/147 msec) when the listeners need help but longer duration (81 msec vs. 65/65 msec) when the listeners can guess the word from context. Pearson's correlation between conditional probability and duration is significant for all three groups ($p < 0.00$).

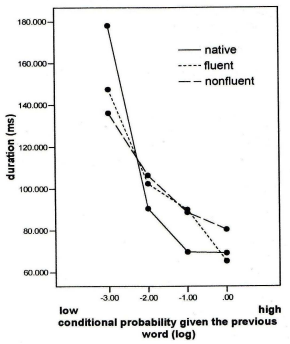

Figure 4: Duration vs. conditional probability for native, fluent, and non-fluent speakers. To the left, the word is relatively surprising, to the right it is expected and predictable.

The same patterns were found differentiating laboratory speech and conversational speech by comparing tonal production of the same words in these two speaking styles [12]. Tonal reduction was stronger in conversational speech but only in position where the information is predictable, as in the middle of long words.

6. CONCLUSION

In this paper, it is argued that much of what is described as "variation" is actually lawful. Some syllables are articulated more carefully than others with the goal of more reliable transmission of information. This care of articulation can be expressed as closer adherence to articulatory targets (such as greater weights in Stem-ML) and/or longer duration. This form of variation is most important in conversational and fluent speech.

Laboratory speech and less-fluent speech has properties that are more similar to Figure 3 (a), where syllables, tones, or words are produced with comparable weights. The speech sounds monotonous and drastic reduction is rare. Conversational speech and fluent speech contain cases that are similar to Figure 3 (b) and (c), where speakers vary the way syllables, tones and words are produced to achieve higher efficiency in communication. Under this view, the difference of laboratory speech, spontaneous speech, fluent speech and non-fluent speech can be explained, and it is clear why evidence of pre-planning shows up more frequently in spontaneous speech and fluent speech.

7. ACKNOWLEDGEMENT

This project is based upon work supported by the National Science Foundation under Grant numbers VIS-0623805 (2007–2010), IIS-0534133 (2006–2009), IIS-0414117 (2004–2007) and a Critical Research Initiative Grant from the University of Illinois at Urbana-Champaign (2005–2008).

8. REFERENCES

[1] Bell-Berti, F., Harris, K. S. 1979. Anticipatory coarticulation: Some implications from a study of lip rounding. *JASA* 65, 1268–1270.

[2] Browman, C. P., Goldstein, L. 1992. Articulatory phonology: An overview. *Phonetica* 49, 155–180.

[3] Kochanski, G., Shih, C. 2003. Prosody modeling with soft templates. *Speech Communication* 39(3-4), 311–352.

[4] Kochanski, G., Shih, C. 2003. Quantitative measurement of prosodic strength in Mandarin. *Speech Communication* 41(4), 625–645.

[5] Kochanski, G., Shih, C. 2006. Planning compensates for the mechanical limitations of articulation. *Proceedings of Speech Prosody 2006* Dresden, Germany.

[6] Li, J., Zheng, F., Xiong, X.-Y., Wu, W. 2003. Construction of large-scale Shanghai Putonghua speech corpus for Chinese speech recognition. *Proceedings of the Oriental-COCOSDA* Sentosa, Singapore. 62–69.

[7] Lindblom, B. 1990. Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastel, W. J., Marchal, A., (eds), *Speech Production and Speech Modelling*. Dordrecht: Kluwer 403–439.

[8] Livescu, K. *Feature-Based Pronunciation Modeling for Automatic Speech Recognition*. PhD thesis MIT.

[9] Öhman, S. E. G. 1966. Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America* 39, 151–168.

[10] Pierrehumbert, J. 1980. *The Phonology and Phonetics of English Intonation*. PhD thesis MIT.

[11] Pierrehumbert, J. 1990. Phonological and phonetic representation. *Journal of Phonetics* 18, 375–394.

[12] Shih, C. 2005. Understanding phonology by phonetic implementation. *Proceedings of Interspeech 2005* Lisbon, Portugal.

[13] West, P. 2000. Perception of distributed coarticulatory properties of English /l/ and /r/. *Journal of Phonetics* 27, 405–425.

[14] Xu, Y. Speech as articulatorily encoded communicative functions. *Proceedings of ICPhS 2007* Germany.

[15] Xu, Y. 2005. Speech melody as articulatorily implemented communicative functions. *Speech Communication* 46, 220–251.

[16] Yoon, S.-Y. 2007. Word probability in L2 learner's speech production– Putonghua retroflex consonant spoken by Shanhainese speakers. University of Illinois at Urbana-Champaign.