Modeling Intonation: Asking for Confirmation in English

Chilin Shih[*] and **Greg Kochanski**[†]

[*] Independent Researcher
[†] Rutgers University, Department of Chemistry and Chemical Biology

ABSTRACT

In this paper, we investigate how intonation is used to confirm a word in English. This intonation type is challenging to model, as it mixes narrow focus and question with variations based on accent location, phrasing and speaking rate.

We build a model that predicts the intonation from the text, using an extremely simple intonational phonology. One can interpret some of the parameters of the model as detailed description of accent shapes and others as prosodic strengths which carry phrasing information. The RMS deviation is 21 Hz or 1.7 semitones, a result comparable to machine learning methods, but with far fewer parameters that need to be learned.

Furthermore, the model handles both fast and slow speech with the same set of parameters in a principled way. The model incorporates some aspects of muscle dynamics, and its ability to predict f_0 at different speaking rates is confirmation that an articulatory approach to f_0 modeling is appropriate.

1 INTRODUCTION

What is your phone number?

301-493-1212.

I am sorry. Is it 301-4**9**3-1212?

Often, in a dialog, a speaker is confident of most of the information in a long list except for one digit in a telephone number, one letter in the confirmation code, or one topping on the pizza. Using intonation is the most natural and effective way to elicit confirmation in this situation.

The speaker draws attention to the word in question by putting a narrow focus on it. The rest of the information becomes a background, guiding the listener to the problematic area. Proper modeling of this intonation is particularly useful in human-machine interaction.

In this paper, we investigate how intonation is used to confirm a word in English. Our primary data are digit strings, as above, where the subject is told to ask for confirmation of a digit (above: the fifth digit, **9**). This intonation type is a challenge to model, because the focus interacts with phrasing, the overall question intonation, and speaking rate.

Our data raises many questions: How to model accents that ride on a high pitch plateau? How to model the interaction of two rising gestures at different distances apart? Is it appropriate to use different phonological representations for fast and slow speaking rates?

We solved many of the problems by building a Stem-ML intonation model [1]. Our model predicts the intonation from the text with very few parameters that need to be learned. One can interpret some of the parameters of the model as detailed description of accent shapes and some others as prosodic strengths which carry phrasing information.

The model handles fast and slow speech in a principled way with the same set of parameters. We believe that the speech-rate dependent f_0 pattern is a consequence of muscle dynamics, and that future modeling of f_0 should incorporate this aspect.

2 DATA

In order to test models of this type of intonation, we collected a corpus of speech from a female, native-born, professional speaker. The database consists of 200 digit sequences, organized in 16 blocks, with variations in phrasing, speaking rate, and which digit in the sentence was to be confirmed. Sentence order was randomized inside each block. Nine of the sentences were dropped in this initial study: three for "oh" *vs.* "zero", and six because pauses were inserted after the focus.

Phrasing: There are two types of phrasing: 12 digit sequences simulating (shortened) credit card numbers (*e.g.* 9478-1509-7091) and 10 digit sequences simulating telephone numbers (*e.g.* 301-123-5045).

Speed: The credit card numbers were read slowly; one set of telephone sequences were read slowly and the other sets fast.

Reading instructions: The speaker was presented with dash-separated digit strings. She was asked to group the digits into credit card style or phone number style, but not to pause at the dash. Each block con-

ISBN 1-876346-48-5

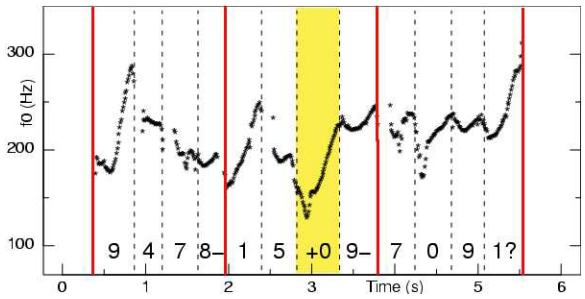

Figure 1: *Phrasing is marked by initial high pitch in the pre-focus region of a confirmation question.*

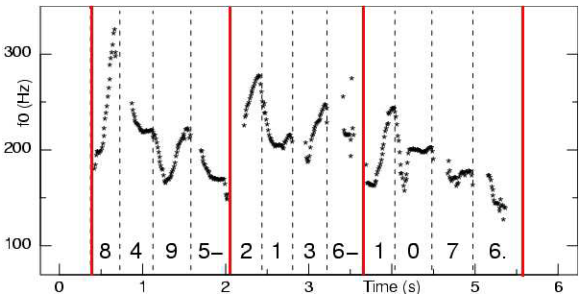

Figure 2: *Phrasing is marked by initial high pitch in declarative sentences.*

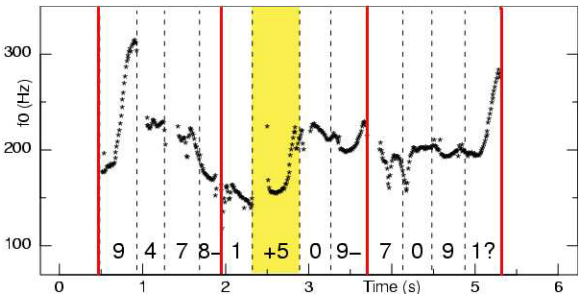

Figure 3: *Phrasing interact with focus. A phrase initial digit immediately preceding a focus is not marked with high pitch.*

sisted of one string read as declarative sentence, one as a yes/no question, and 10 (or 12) asking for confirmation on different digits: *You know you've got most of the numbers but are not sure about the one written in red [boldface]. You are trying to confirm whether this digit is correct.* 901-1**0**9-9091?

3 OBSERVATIONS

The intonation pattern of a confirmation question is realized consistently in our data. Figure 1 is a typical example. The speaker put narrow focus (strong emphasis) on the digit she tried to confirm. The narrow focus is used in the context of a question, and the observed intonation pattern is a combination of these two functions. There is a final rise at the end of the sentence, as in a typical English yes/no question. In addition, The speaker used a strong rising accent on the focal digit. The pitch remains high after the focus. When the focus is close to the end, the confirmation rise and the final rise fuse together.

All figures in this section display f_0 tracks in Hertz as a function of time (seconds). Vertical dashed lines mark word boundaries. Dash "-" marks phrasing as indicated in the text. The intended phrasing is also marked in the figures with a thick (or red in color display) solid line. There may or may not be acoustic correlates at the indicated phrasing boundaries. A leading plus "+" sign marks the digit to be confirmed, which is also shaded (in yellow).

3.1 PHRASING

Phrasing as indicated by the dash in text is clearly marked in the prosody: Pitch rises on the phrase initial digit and falls towards the end of the phrase. This is found in all declarative sentences, such as Figure 2. The pre-focus pitch contour and phrasing are similar to that in the comparable region of a declarative sentence, as can be seen in the first two phrases of Figure 1.

Further examination shows that phrasing effect interacts with focus. In sharp contrast to Figures 2 and 1, the phrase initial digit *one* in Figure 3 is not marked with high pitch. The speaker seems to employ a rhythmic consideration here, de-accenting the phrase-initial digit to avoid putting two strong accents too close together. This interaction of focus and phrasing cannot be modeled using a linear additive model for pitch, but our model can explain it as a consistent reduction in strength before the focal digit.

3.2 SPEAKING RATE

Having parallel data at two speech rates raises questions about the proper phonological representation of these contours. The surface contours are different enough to suggest different representations, but is the phonology really different? Are people actually shifting grammars as they change speaking rate?

We will first evaluate whether all words after the narrow focus are de-accented [2]. Our fast speech data in Figure 4 matches this description reasonably well: other than the final rise, pitch movements after the narrow focus are suppressed. Following the ToBI transcription system [3, 4], the focus and the post-focus area in Figure 4 may be represented phonologically as *L*+H H- H%*, a rising accent (L*+H) on the focal digit, a high phrase tone (H-) accounting for the high plateau after the focus, and a high boundary tone (H%) accounting for the final rise.

In contrast, the slow speech data in Figure 5 shows that accents are present on each word after the focus. Phonologically, this requires an accent on every digit.

ISBN 1-876346-48-5

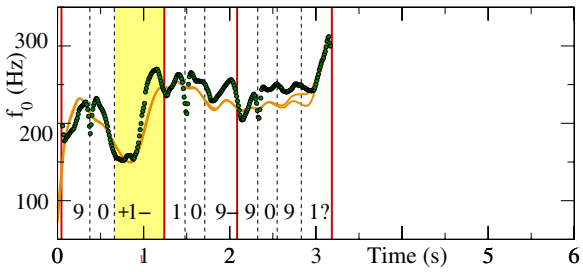

Figure 4: *Fast speech: Accent movements are less obvious after the narrow focus. F_0 measurements are the dark circles; model predictions are shown as the solid orange (gray) lines.*

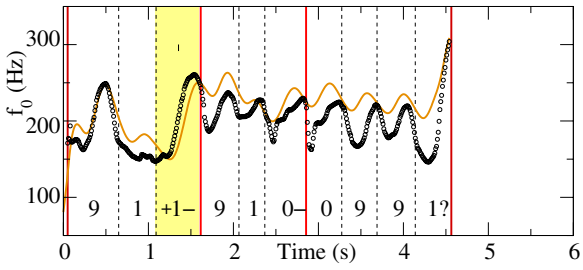

Figure 5: *Slow speech: Accents are present after narrow focus. Orange (grey) curve is the model prediction from the training set onto the test set.*

By incorporating a muscle dynamic model into intonation modeling, we show that the surface difference in fast and slow speech is the natural result of timing variation. There is no need to assume different phonological representations.

4 MODEL

The best way to establish properties of a language is to capture those properties in a quantitative model, then to test the model. Consequently, we built a model of intonation that uses Stem-ML [1][5]. Stem-ML simulates the dynamics of the relevant articulators (for intonation, these are the vocalis and crycothyroid muscles, primarily) by assuming that the speaker is striking a balance between the effort used in speech and the probability that his/her speech will be misinterpreted. This balance is embodied in the Stem-ML *strength* parameters, which, for a given word, measure how important it is that the articulator motions be executed accurately.

In normal English text, one might need to allow the strength of each word to vary independently, because the importance of a word may well be affected by semantics and syntax. However, all digits have equivalent semantic and syntactic features, so the strength of a digit should be determined only by its position in the phrase and its position in the utterance.

This equivalence allowed us to use the simple model for

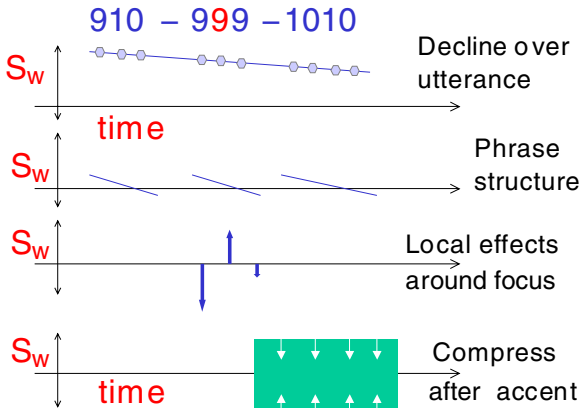

Figure 6: Components used in calculating the log(*Strength*) of words in the model.

the word strength, shown schematically in Figure 6. The log of the strength is given by a linear decline through each 3- or 4-digit phrase, added to a linear decline throughout the utterance. Both rates of decline are adjustable parameters in the model which are fit to the data. Then, the strengths of the focal word and its immediate neighbors are changed (the magnitude of each of the three changes is likewise a parameter). Finally, the strengths after the focus are scaled and shifted (two more adjustable parameters).

Stem-ML also has templates, which we associate with words, and which influence the shape of f_0 *vs.* time in the vicinity of the word. If a word has a relatively high strength, f_0 will locally match the word's template. If the word has a low strength, the tone shape will be controlled by the surroundings.

We assume that all the words outside the sentence focus share the same template. Then, we add a special template for the focal word, one template for an initial boundary tone, and one for a final boundary tone. The boundary tone templates overlap the templates for the initial and final words. Our model does not, in any way, specify the shapes of the templates; they are derived by matching the model to the corpus of data. The sharing of templates specifies which parts of the utterance are linguistically equivalent. If they truly *are* equivalent, then the model may be able to reproduce the observed intonation. If they *aren't* equivalent, for instance if some boundary tones were different from others, then the model will not be able to capture the difference.

The model has 48 adjustable parameters. All of the parameters are global parameters which are shared among all sentences; that is, we left no room in the model for sentence-specific variation. We fit the model to the 43 sentences that are (a) composed of voiced digits one, nine and zero, (b) have no pauses, and (c) were not declarative. The model's parameter density is 1.1 parameters per sentence; it is more compact and

ISBN 1-876346-48-5

makes stronger predictions than our earlier work on Mandarin read speech [5], because the speech in this work is more stylized and is from a limited domain.

We fit the data using techniques similar to [5], except that we used a generalized Jackknife [6] procedure to estimate the uncertainties in our fitted parameters. To do this, in each run, we assigned utterances to a test set with a 10% probability, then fit a model to the remaining training set. We followed this procedure 22 times, using the best-fit parameters from one model to initialize the fitting procedure for the next. Statistical confidence intervals can then be deduced from the standard deviation of the 22 best-fit values for that parameter.

Examples of the model's predictive ability are shown in Figures 4 and 5. The dark circles are the measured f_0, and instances of the model are shown as orange(gray) solid lines. The data shown in those figures was not used to compute the solid lines: they are predictions, based on other utterances in the corpus. The two curves in Figure 4 are generated from different training sets, and show that the results of the model are quite consistent as the training data is changed.

5 RESULTS

The RMS deviation is 0.21 Barks, which corresponds to approximately 21 Hz or 1.7 semitones. The result is surprisingly good especially considering how few parameters are being used.

The model captures the slow and fast speech variations naturally without any need to adjust the model for fast or slow speech, or any parameter addressing this aspect of the variation. Slow speech has more pitch movement while fast speech has relatively smooth pitch. While many intonation schemes would require a categorically different set of accents to express the difference between fast and slow speech, our model provides a unified view, generating both variants from the same phonological symbols. This property may lead to a significant simplification of intonation phonology.

Some of the parameters of the model have unambiguous interpretations. We will now discuss these.

Boundary Tones: We find that the strength and length of the initial/final boundary tone is almost independent of the strength or duration of the corresponding word. This is consistent with the boundary tone being a property of the sentence as a whole. Similarly, boundary tone lengths are nearly constant, independent of the durations of the corresponding words. Final boundary tones are significantly longer than initial boundary tones (99 ± 9 ms[1] *vs.* 28 ± 13 ms). Together, these results provide strong support that a boundary tone should be part of a description of English intonation.

Strength: The phrase-scope and utterance-scope strength variations are reduced after the focal digit. The strength of the digit just before the focal digit is drastically reduced, to 33% or less[2] of what it would have been in the absence of focus.

Accent Shapes: The shape of the accent on the focal digit (a step) is dramatically different from the shape on all other digits. Accent lengths are roughly proportional to the length of the word.

6 CONCLUSION

We describe a Stem-ML model of a small domain of English intonation that is derived from an extremely simple intonational phonology, yet still accurately captures the speaker's pitch contour. The success of this model raises the possibility that much of the complexity of current phonological theories arise because they are attempting to describe phenomena that are really phonetics. For instance, we suggest that the fast and slow speech phenomena are best explained in phonetic variations, rather than phonological representations. The apparent change in phrasal structure when the focal digit is second in a phrase can be explained more simply by a uniform reduction in the strength of the digit before the focal one.

REFERENCES

[1] Greg Kochanski and Chilin Shih, "Prosody modeling with soft templates," *Speech Communication*, vol. 39, no. 3-4, pp. 311–352, February 2003.

[2] Elisabeth O. Selkirk, *Phonology and Syntax: The Relation between Sound and Structure*, The MIT Press, Cambridge, MA, 1984.

[3] Janet Pierrehumbert, *The Phonology and Phonetics of English Intonation*, Ph.D. thesis, MIT, 1980.

[4] M. E. Beckman and G. Ayers, "Guidelines for ToBI labeling," 1997, http://www.ling.ohio-state.edu/phonetics/ToBI/ToBI.0.html.

[5] Greg Kochanski, Chilin Shih, and Hongyan Jing, "Hierarchical structure and word strength prediction of Mandarin prosody," *International Journal of Speech Technology*, vol. 6, no. 1, pp. 33–43, January 2003.

[6] Bradley Efron, *The Jackknife, the Bootstrap, and other resampling plans*, Number 38 in CBMS-NSF Regional Conf. Series in Applied Mathematics. SIAM, 1982.

[1] 1-σ error bars

[2] one-sided, 95% confidence interval

ISBN 1-876346-48-5