A Generative Phonetic Analysis of the timing of L- Phrase Accents in English

Edward Flemming

Massachusetts Institute of Technology (USA)
flemming@mit.edu

This paper builds on Pierrehumbert's foundational work on the generative phonetics of English intonation [1] through an analysis of the timing of the English low phrase accent (L-) in H*L-L% and H*L-H% melodies, developing a model of the realization of tones that has three key elements:

(i) F0 trajectories are modelled as the response of a dynamical system to a control signal that consists of a sequence of step functions connected by linear ramps.

(ii) The mapping from tones to F0 trajectory is derived by optimization with respect to conflicting constraints.

(iii) Variation in realization is derived by formulating the grammar as a Maximum Entropy model

Identifying elbows

Studying the timing of L- in H*L- sequences is challenging because L- is realized as an 'elbow' in the F0 trajectory – i.e. a point of inflection rather than a local maximum or minimum – when there is sufficient time between the H* peak and the end of the phrase (fig. 1). Algorithms for locating F0 elbows have been proposed [2, 3], but they are difficult to evaluate because there is no independent evidence concerning the true locations of elbows.

The approach adopted here is to develop a more explicit model of the interpolation between tonal targets. This makes it possible to infer the location of targets by fitting the model to the observed F0 contour. Specifically, the fall from the H* peak is modelled as the response of a cascade of four identical first order linear dynamical systems to a step function input, and the L- plateau is modelled as the response of the same system to a linear input (fig. 2). This type of model has been used to model articulator movements [4], and fits F0 contours much better than the more familiar critically damped second order dynamical systems [5] because it allows for variation in the timing of the acceleration peak.

The timing of L- elbows

This analytical method is used to test two hypotheses concerning the timing of L- elbows: (a) L- occurs at a fixed interval after H*, (b) L- is aligned to the end of the nuclear-accented word [1]. The data were recordings collected for a previous study [6]. The materials consist of 25 two-word phrases read by 15 speakers in a context designed to elicit an H*L-H% melody with H* on the first word (although some utterances were produced with H*L-L%). The number and length of the syllables following the primary stress in the first word were systematically varied (e.g. *á<u>lien</u>* vs. *pá<u>limony</u>*) to test if the interval from the H* peak to L- elbow varies to align L- to the word boundary or not.

The results do not support either hypothesis: L- is not aligned to the word boundary, but there is a significant tendency for L- to occur earlier when the interval between H* and the word boundary is shorter (fig. 3) (cf. [6]). This pattern of realization is analysed as a compromise between two constraints, one enforcing a target duration for the fall from H* to L-, and a second requiring the fall to be completed before the end of the word. The second constraint has lower weight, so the H*-L- duration only decreases by a small proportion of a decrease in duration to the word boundary.

Maximum Entropy phonetic grammar

The variation in L- timing within and between speakers is more striking than the effect of variation due to timing of the word boundary. Variation in phonetic realization can be modelled by adapting the widely used MaxEnt approach to phonological variation [7]. In MaxEnt grammar, candidates are assigned a numerical wellformedness score, H, equal to their summed constraint violations, and the probability of a candidate is proportional to e^{-H}. If the cost of violating a phonetic constraint is proportional to the square of the deviation from the constraint target, then the MaxEnt model derives a normal distribution over the space of realizations, with its mean at the lowest cost realization, and variance dependent on the summed constraint weights.

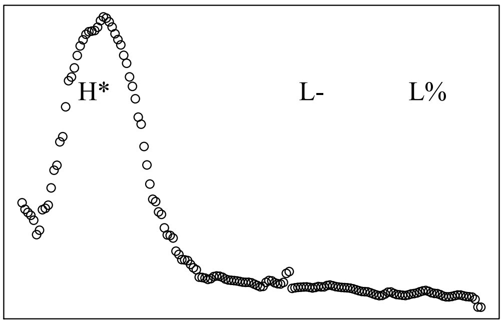

Fig.1 F0 track of an English H*L-L% melody produced on the phrase 'alien annihilator'

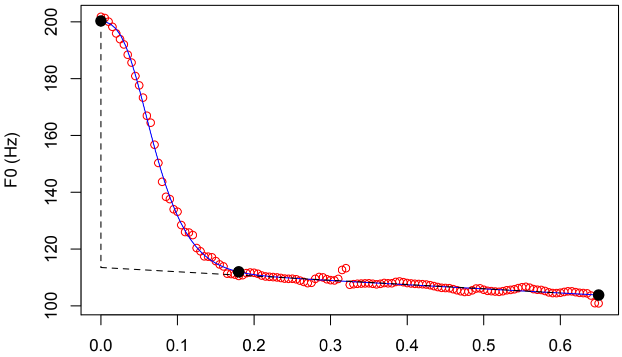

Fig.2 F0 measurements (red), model F0 trajectory (solid line), and input to the production model (dashed lines) for the utterance shown in fig. 1.

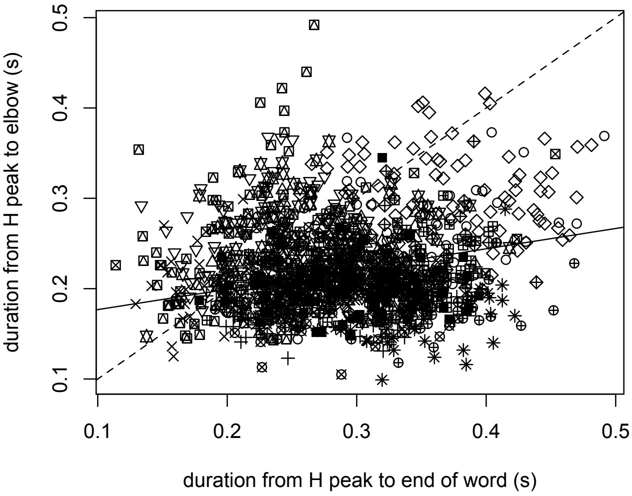

Fig.3 Plot of the relationship between duration from H* to the L- elbow and durationship from H* to end of word. The solid line shows the best-fitting linear relationship, and the dashed line shows the expected trend if L- is aligned to the end of the word. Speakers are plotted with different symbols.

References

[1] Pierrehumbert, J. B. (1980). *The Phonology and Phonetics of English Intonation*. Ph.D. Thesis, MIT, Cambridge, MA.

[2] del Giudice, A., Shosted, R., Davidson, K., Salihie, M., & Arvaniti, A. (2007). Comparing methods for locating pitch "elbows". *Proceedings of the 16th International Congress of Phonetic Sciences*, 1117-1120.

[3] Reichel, U., & Salveste, N. (2015). Pitch elbow detection. *ESSV*, Katholische Universität Eichstätt-Ingolstadt, Germany, 25.–27. March 2015. Ed. Wirsching, G. Dresden: TUD Press, 143−149.

[4] Birkholz, P., Kröger, B.J., & Neuschaefer-Rube, C. (2011). Model-based reproduction of articulatory trajectories for Consonant-Vowel sequences. *IEEE Transactions on Audio, Speech, and Language Processing,* 19, 1422-1433.

[5] Fujisaki, H. & Hirose, K. (1984), Analysis of voice fundamental frequency contours for declarative sentences of Japanese, *Journal of the Acoustical Society of Japan*, 5, 233-242.

[6] Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2010). Turning points, tonal targets, and the English L- phrase accent. *Language and Cognitive Processes,* 25:7-9, 982-1023.

[7] Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. Eriksson, and O. Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, 111–120.