

Categorical effects in production of pitch contours in English

Laura C. Redi

Speech Communication Group, MIT, Cambridge, Massachusetts, U.S.A.

E-mail: redi@mit.edu

ABSTRACT

Previous work suggests that methods used in studying categorical perception for segmental contrasts may be useful in determining the representation of suprasegmental contrasts, particularly those based on intonational characteristics. This experiment investigated phonological representations for intonation contours in English by determining the extent of categorical effects in a task involving imitation of synthetic speech continua in which a fundamental frequency maximum (peak) or minimum (valley) was shifted in 25 ms increments through a target two-syllable sequence with a particular stress pattern (WS or SW). Imitations of both WS and SW peak continua showed categorical effects consistent with two categories of representation, while imitations of valley continua were inconclusive. Results are interpreted with respect to phonological categories of representation that have been proposed for English intonation.

1. INTRODUCTION

It is now well-established that the perception of acoustic continua is often mediated by phonological categories of representation. This result, termed *categorical perception*, has most frequently been investigated using acoustic continua related to established segmental contrasts. (See e.g. Repp [1] for a review.) However, categorical effects have also been demonstrated in relation to *suprasegmental* contrasts, in particular for continua involving fundamental frequency (F0) gradation [2,3]. This suggests that investigative techniques utilized in studying segmental contrasts may be helpful in determining the nature of the representation for suprasegmental aspects of language, particularly intonation. This paper examines the nature and extent of categorical effects in the imitation of synthetic stimuli in which F0 has been varied along a continuum, in order to empirically evaluate aspects of one proposed phonological theory of English intonation, the autosegmental-metrical (AM) model [4,5].

A consistent finding that has emerged from studies demonstrating categorical effects for F0 continua is the importance of the timing of certain characteristics, such as local maxima, rises, or falls, with respect to the segmental string [3,6,7]. Of these characteristics, the timing of F0 maxima and minima (or *peaks* and *valleys*) relative to the segmental string, which is termed *alignment*, is of especial interest, for two reasons. First, a number of phonetic studies have shown that the timing of F0 peaks and valleys relative

to the string of segments is remarkably stable within a variety of languages, and differential alignment of these F0 attributes in some languages is a consistent phonetic correlate of phonological distinctions [8,9,10]. Moreover, listeners are sensitive to the timing of such F0 characteristics and can use this information to distinguish lexical items or semantic intent [2,11].

These findings suggest the possibility that phonetic differences in F0 peak and/or valley timing may give rise to phonological categories cross-linguistically, just as voice-onset time (VOT) variation may give rise to one or more distinctions (e.g., [+voiced] or [-voiced]). This possibility motivates close examination of phonological models of intonation and empirical testing of the proposed mappings between categories of representation and the phonetic alignment of F0 peaks and valleys.

The experiment reported here tests some predictions of the AM model of intonation regarding the proposed mapping between phonological categories of representation and F0 peak or valley alignment. This is accomplished by assessing the nature and extent of categorical effects in an imitation task when F0 peak and valley timing are varied along an acoustic continuum. The AM model makes several specific predictions regarding whether shifting an F0 peak or valley through a weak-strong (WS) or strong-weak (SW) syllable sequence will give rise to one versus two phonological categories. The predictions of the model are discussed in the following section.

2. METHOD

2.1 STIMULI

Stimuli consisted of synthetic speech materials in which an F0 peak or valley was shifted in 25 ms increments across a target two-syllable WS or SW sequence, for a total of four acoustic continua. Each continuum was based on a phrase with controlled segmental composition. Target WS or SW sequences contained exclusively voiced, mostly sonorant segments, with a nasal-consonant (or consonant-nasal) sequence to facilitate spectral identification of the syllable boundary. Furthermore, both target syllables had either high or non-high vowels to minimize possible effects of intrinsic F0 on peak or valley location. The overall intonation pattern for each phrase was selected so that the target two-syllable sequence was either high in the pitch range (in the case of F0 peak continua) or low in the pitch range (in the case of F0 valley continua), to facilitate testing of the AM model's predictions. The phrases selected for F0

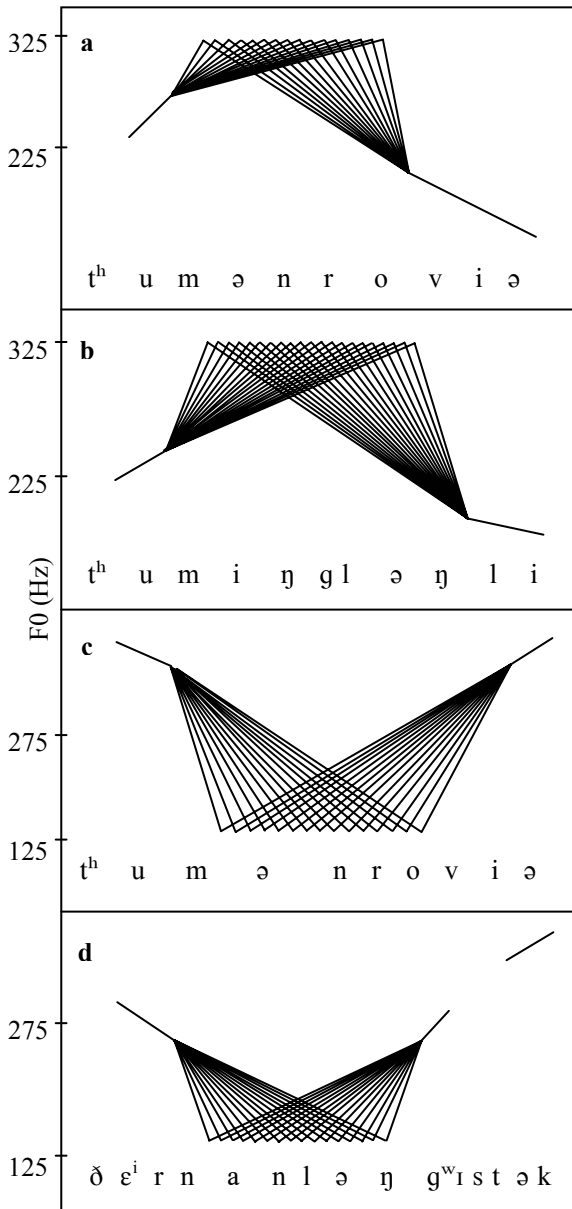


Figure 1: Stimulus continua used in the production experiment. Top to bottom: WS Peak, SW Peak, WS Valley, and SW Valley continua.

peak continua were *To Monrovia*, which contains the WS sequence [mən'ro], and *Too minglingly* (from the SW verb *mingle*), which contains the SW sequence ['miŋgləŋ]. The phrases selected for F0 valley continua were *To Monrovia?*, which again contains the WS sequence [mən'ro], and *They're nonlinguistic?*, which contains the SW sequence ['nanləŋ].

Each phrase was recorded in a sound-attenuated room and digitized at 22 kHz using a DAT recorder. F0 continua were created using the Praat speech analysis and synthesis software package [12], with stylized F0 contours consisting of straight line segments. A total of 16, 21, 15, and 16 stimuli were created for the WS Peak, SW Peak, WS Valley,

and SW Valley continua, respectively (Figure 1). Each of the four continua was predicted to correspond to either one or two phonological categories (termed *pitch accents* in the AM model), as follows:

1) *WS Peak continuum* (Figure 1a): The leftmost member of this continuum corresponds to the H+L* pitch accent (H+!H* in the ToBI system of intonational notation [13]), while the rightmost member corresponds to a H* accent.

2) *SW Peak continuum* (Figure 1b): Both ends of this continuum are analyzed as a H* accent, since for this accent an F0 peak is assumed to be aligned either with a stressed syllable or the following weak syllable [6,13].

3) *WS Valley continuum* (Figure 1c): The leftmost member of the continuum corresponds to a L+H* accent, while the rightmost member is an example of a L* accent.

4) *SW Valley continuum* (Figure 1d): The leftmost member of the continuum corresponds to a L* accent (on ['nan]), while the rightmost member corresponds to a L+H* accent (on [gwis]).

2.2 SUBJECTS AND TASK

Subjects were 21 native English speakers who were students or staff at MIT or a nearby college. All subjects reported normal hearing.

Stimuli were presented over headphones to subjects at a comfortable volume in a sound-attenuated room. The subjects were instructed to imitate each phrase that they heard as closely as possible in a comfortable pitch range. The text of each phrase appeared on a computer screen. Stimuli from a given continuum were presented in blocks, where the order of presentation with each block was randomized. The stimuli for each continuum were presented in three separate blocks, for a total of twelve blocks. Each block of trials was preceded by practice trials consisting of a subset of stimuli drawn from the upcoming block. The subjects' productions were digitized at 16kHz in real time using MARSHA software by Mark Tiede.

2.3 ANALYSIS

The temporal location of the F0 peak or valley in subjects' productions was determined automatically using Praat and checked by hand for accuracy. If the F0 peak or valley did not occur within the time spanned by the target SW or WS syllable sequence, the trial was discarded ($n < 10$). In the event that segmental effects on the F0 contour resulting from transient pressure buildup induced a localized F0 maximum or minimum, the next highest max or min was taken as the location of the peak or valley, respectively.

In addition to determining the time of the peak or valley, the time of the onset of the first syllable and both the onset and offset of the second syllable in the SW and WS syllable sequence was determined. The peak and valley location was then normalized relative to these temporal positions, using the formula given in equation (1). In this formula, t is the time of the peak or valley, t_0 is the start of the first

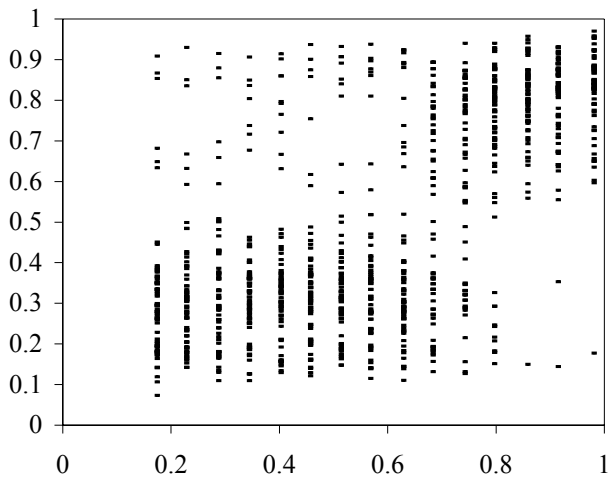


Figure 2: Normalized peak locations for WS Peak stimuli and subjects' productions.

syllable in the target SW or WS syllable sequence, and d_1 and d_2 are the durations of the first and second syllables, respectively, in the target syllable sequence. Thus, the normalized peak (or valley) location ranged from 0 to 1.0.

$$\text{Normalized peak (valley) location} = \frac{t - t_0}{d_1 + d_2} \quad (1)$$

During analysis, it was discovered that the final stimulus in the WS Peak series had a minimum just after the end of the target strong syllable. Since the normalized peak location in this stimulus was therefore greater than 1.0, productions in response to this stimulus were discarded from analysis.

3. RESULTS

3.1 F0 PEAK STIMULUS CONTINUA

Figures 2 and 3 show normalized peak locations for stimuli in the WS Peak and SW Peak continua, respectively, plotted against normalized peak locations in subjects' imitations. Figure 2 shows that the normalized peak locations in most imitations of stimuli 1-9 fall into the range 0.1-0.45, while normalized peak locations in most imitations of stimuli 13-15 fall within the range 0.55-0.95, with a shift occurring around stimuli 10-12. Similarly, Figure 3 shows that the normalized peak locations for most imitations of stimuli 1-11 fall into the range 0.2-0.5, while normalized peak locations for most imitations of stimuli 16-21 fall into the range 0.6-0.9, with a shift occurring around stimuli 12-15.

3.2 F0 VALLEY STIMULUS CONTINUA

Figures 4 and 5 show the normalized valley locations for stimuli in the WS Valley and SW Valley continua, respectively, plotted against the normalized valley locations in subjects' imitations. Neither figure shows a shift in preferred location for the alignment of the F0 valley. In general, normalized peak locations for subjects' productions fall within the range 0.2-0.9 for all stimuli in both the WS Valley series (Figure 4) and the SW Valley

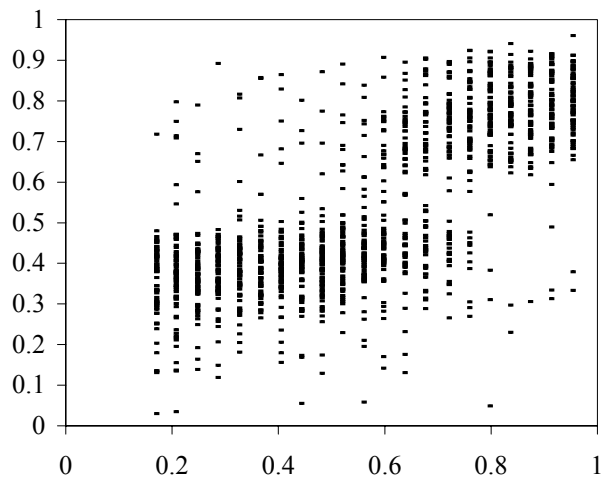


Figure 3: Normalized peak locations for SW Peak stimuli and subjects' productions.

series (Figure 5). Inspection of individual data showed variability across subjects in F0 valley alignment.

4. DISCUSSION

Because acoustic continua based on F0 gradation have been shown to sometimes elicit categorical effects [2,3], methods used in investigating categorical perception may be used to evaluate models of the phonological representation of intonation. The findings presented here are interpreted with respect to phonological categories proposed by the AM model of English intonation [4,5].

The shift in values for normalized peak location observed in response to WS Peak stimuli (Figure 2) suggests the presence of a category boundary along this continuum. These data therefore support the distinction proposed by the AM model between the H+L* (or H+!H*) accent and the H* accent categories. This data represents the first empirical support of this proposed accent distinction.

Moreover, the shift in normalized peak values observed in response to SW Peak stimuli (Figure 3) suggests the presence of a category boundary along this continuum. This data is therefore inconsistent with the single phonological category proposed by the AM model (H*) and instead suggests the existence of two separate phonological categories along this dimension.

A possible alternative explanation for the categorical effect seen in Figure 3 which might be offered is that the SW Peak continuum was interpreted by subjects as corresponding to two intonational categories proposed under the AM model, namely L+H* vs. L*+H. This distinction is assumed to give rise to a difference in F0 peak timing: the peak occurs on the stressed syllable for L+H* and after the stressed syllable for L*+H [3]. Because these accents each involve a L tone, this explanation predicts that when there is a peak in the weak syllable of the SW sequence, subjects will produce a locally low F0 value on the stressed syllable. To check this possibility, the average F0 on the stressed

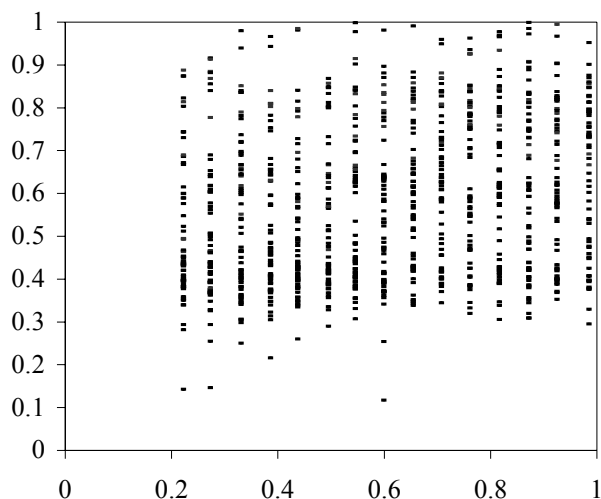


Figure 4: Normalized valley locations for WS Valley stimuli and subjects' productions.

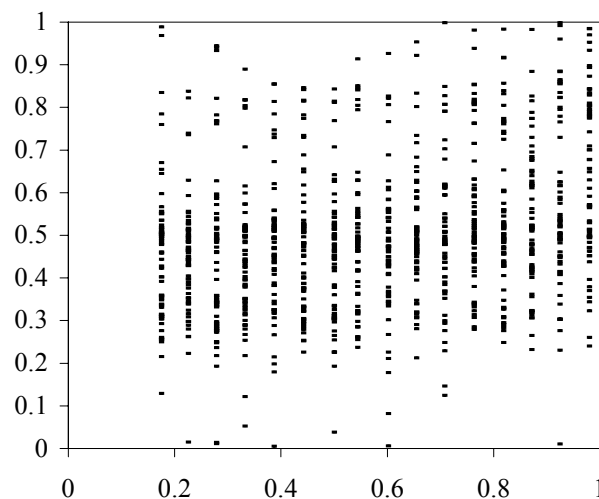


Figure 5: Normalized valley locations for SW Valley stimuli and subjects' productions.

syllable ([miŋ]) was compared with the average F0 of the vowel of the preceding syllable ([t^hu]) for productions in which the peak was aligned with the weak syllable. In 92% of cases, the stressed syllable had a higher average F0 than the /u/ of [t^hu], suggesting that subjects were not hearing and reproducing the L+H* vs. L*+H distinction.

The variability in subject responses to valley stimuli meant that it was not possible to determine the extent of support for the remaining proposed accent distinctions. The lack of consistency in responses to these stimuli may be due to any of several factors. One possibility is that F0 valleys may be harder to perceive. Another possibility is that subjects had difficulty reproducing stimuli which were low in their pitch ranges. Further work will be needed to decide among these and other possible explanations.

4. CONCLUSIONS

This paper presents evidence of categorical effects for acoustic continua in which an F0 peak is shifted through a WS or SW syllable sequence. The first empirical support is presented for the claimed distinction between H+L* (or H+IH*) accent and the H* accent proposed by the AM model of English intonation. In addition, evidence is presented which suggests the need for revision of the phonological category of H* proposed by the AM model.

ACKNOWLEDGEMENTS

This work was supported by a NSF Graduate Fellowship and a NIH Training Grant to the Harvard-MIT Speech and Hearing Bioscience and Technology Program. I wish to thank Stefanie Shattuck-Hufnagel for helpful feedback on this paper.

REFERENCES

[1] B.H. Repp, "Categorical perception: Issues, methods, findings," in *Speech and Language: Advances in Basic*

Research and Practice, vol. 10, N.J. Lass, Ed., pp. 243–335. Orlando FL: Academic Press, 1984.

[2] K.J. Kohler, "Categorical pitch perception," in *Proceedings of the 11th ICPhS*, Tallinn, Estonia, vol. 5, U. Viks, Ed., pp. 331–333, 1987.

[3] J.B. Pierrehumbert and S.A. Steele, "Categories of tonal alignment in English," *Phonetica*, vol. 46, pp. 181–196, 1989.

[4] J.B. Pierrehumbert, "The phonology and phonetics of English intonation," Ph.D. thesis, MIT, 1980.

[5] M.E. Beckman and J.B. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology Yearbook*, vol. 3, pp. 15–70, 1986.

[6] J. Verhoeven, "The discrimination of pitch movement alignment in Dutch," *Journal of Phonetics*, vol. 22, pp. 65–85, 1994.

[7] T. Rietveld and C. Gussenhoven, "Aligning pitch targets in speech synthesis: effects of syllable structure," *Journal of Phonetics*, vol. 23, pp. 375–385, 1995.

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

[9] D.R. Ladd, *Intonational Phonology*, Cambridge University Press, 1996.

[10] A. Arvaniti, D.R. Ladd, and I. Mennen, "Stability of tonal alignment: the case of Greek prenuclear accents," *Journal of Phonetics*, vol. 26, pp. 3–25, 1998.

[11] G. Bruce, *Swedish word accents in sentence perspective*, Lund: Gleerup, 1977.

[12] P. Boersma, "Praat, a system for doing phonetics by computer," *Glott International*, vol. 5, no. 9/10, 2001.

[13] M. Beckman and G. Ayers-Elam, "Guidelines for ToBI labeling," www.ling.ohio-state.edu/research/phonetics/E_ToBI, 1997.