DO WORDS SING THEIR OWN TUNES? WORD-SPECIFIC PITCH REALIZATIONS IN MANDARIN AND ENGLISH

Yu-Ying Chuang[1], R. Harald Baayen[1], and Melanie J. Bell[2]

University of Tübingen[1]; Anglia Ruskin University[2]
yu-ying.chuang@uni-tuebingen.de; harald.baayen@uni-tuebingen.de; melanie.bell@aru.ac.uk

ABSTRACT

This study investigates the pitch realizations of disyllabic words in Mandarin and English spontaneous speech. From two spoken corpora, we obtained F0 measurements for Mandarin words with the Rise-Fall tonal pattern, and for English words with left stress. Analyses were conducted on the pitch contours of these words, using Generalized Additive Mixed Models (GAMMs). Pitch excursions in Mandarin were generally larger, and had narrower confidence intervals compared to English, consistent with Mandarin being a tone language. Surprisingly, for both languages, word identity accounts for a considerable amount of variance in F0, independently of other covariates such as speaker sex and speech rate. The GAMMs reveal that words have their own characteristic pitch contours, just as they have their own characteristic segmental realizations. A word's specific F0 contour probably reflects the contexts and context-sensitive senses that are most common and characteristic for that word.

Keywords: Word-specific pitch, Mandarin tone, English lexical stress, F0 contour, disyllabic words

1. INTRODUCTION

In language, pitch is multi-functional. Pitch can reflect speakers' emotions and direct listeners' attention to the important parts of utterances.[1] At the lexical level, in tone languages, it serves to distinguish word meanings, and in non-tone languages, it contributes to perceived stress. However, speakers produce far more variation in pitch than would be predicted by discrete categories such as tone or stress [1, 2]. Various theories have been put forward to account for the observed pitch variation in laboratory speech, and many factors have been shown to affect pitch. Nevertheless, the way in which these factors interact in spontaneous speech production is not well understood.

The present study uses corpus data to investigate the pitch realizations of disyllabic Mandarin and English words in spontaneous speech. We focused on Mandarin words with the Rise-Fall (RF) tonal pattern (e.g., 學校 *xuéxiào* 'school') and English words with initial stressed syllable (left-stressed, e.g., ***hea****vy*). We modeled the pitch contours of our corpus tokens using generalized additive mixed models (GAMM, [3]). GAMMs allow us to model F0 as a non-linear function of time across an utterance, while also including other predictors known to affect pitch, such as speech rate and speaker sex. They can thus capture fine-grained pitch undulations in the realization of words.

Our central question is whether, in spontaneous speech and across languages, pitch contours vary consistently with word. That is to say, do individual word types have their own unique 'pitch signatures' just as they have their individual segmental makeup? In what follows, we first present the analyses and results for Mandarin. We then report our results for English. In the General Discussion, we reflect on the theoretical implications of our findings.

2. MANDARIN

The realizations of Mandarin tones have been widely studied using experimental methods in laboratory settings. It has been found that tonal realizations are subject to a variety of factors, ranging from influences at the local level such as syllable structure and consonant type, to utterance-level effects such as focus or intonation (see [4] for a review). It is therefore usually the case that actual tonal realizations deviate substantially from the canonical contours, and might even completely lose the expected shapes in connected speech [5]. Specifically for the RF pattern, previous studies based on lab speech show that its actual realization is a dipping contour followed by a delayed fall [6]. In an extreme case of severe contraction, neither rising nor falling is preserved [7].

2.1. Data

Instead of using lab-controlled experimentation, the present study investigates the realizations of the

RF contour in spontaneous speech. From the Taiwan Mandarin spontaneous speech corpus [8], we selected 51 RF words, each with at least 20 tokens. After removing tokens with pitch tracking error, we were left with a total of 3860 tokens. For each token, we took F0 measurements every 15 ms. The timepoints of these measurements were subsequently transformed into normalized timepoints from 0 to 1 (TIME). We also calculated the local speech rate for each token (SR), defined as the number of syllables per second within a time window of four words to the right and to the left of the target word. In addition, we coded the tonal context of each token (CONT), i.e., the tone type of the preceding and following words, as well as the token's normalized position in the respective utterance (POS).

2.2. Analyses

We fitted a GAMM to F0, with the following additive components: (1) a main effect of SEX, to take into account that on average, females have higher F0, (2) two TPRS[2] smooths modeling pitch as a function of TIME, one per SEX, (3) two TPRS smooths for SR, one per SEX, (4) a TPRS smooth for POS, (5) two tensor-product interaction smooths for TIME by SR, one per SEX, (6) a tensor-product interaction smooth for TIME by POS, (7) a random-effect factor smooth for TIME by SPEAKER, and (8) a random-effect factor smooth for TIME by CONT:

```
F0 ∼ SEX + s(TIME, by=SEX) +
     s(SR, by=SEX) + s(POS) +
     ti(TIME, SR, by=SEX) +
     ti(TIME, POS) +
     s(TIME, SPEAKER, bs="fs", m=1) +
     s(TIME, CONT, bs="fs", m=1)
```

To address our research question, we added to these control variables our main variable of interest, namely a factor smooth for word type (WORD). With this variable included, the addition of further word-based control variables such as consonant types, vowel types and syllable structure invariably led to unacceptably high collinearity, rendering model parameters difficult to interpret [9]. We therefore excluded these additional variables from the analysis.

2.3. Results

Figure 1 presents the partial effects of the main-effect smooths modeling F0 as a function of TIME, for females (left) and males (right). Similar to results based on lab speech, the contours in spontaneous speech are characterized by an initial shallow fall, followed by a long rise, and finally a fall. The partial effect for males has somewhat reduced minima and maxima, except for the final fall, which ends lower. These contours are fine-tuned by the other control variables in the baseline model.

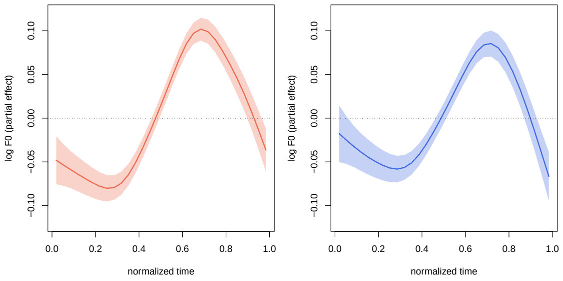

Figure 1: Partial effects of the smooths for F0 as a function of normalized time for females (left) and males (right).

Inclusion of the by-WORD factor smooth improved the model fit by no less than 10,250 AIC units. This result implies that how an RF pitch contour is realized is co-determined by the identity of the word. To ensure that this improvement was not an artefact, we randomized the word labels, fitted GAMMs with and without WORD as predictor, recorded the difference in AIC, and repeated this process 100 times to produce a distribution of model fit improvements in AIC units. The mean of this distribution was 1443, and the SD was 224. It is therefore very unlikely that the improvement of model fit with true word labels (10,250 units) is a statistical artifact. Spontaneously produced RF words in Mandarin thus emerge as having individual pitch signatures, just as they have their own segmental signatures.

Figure 2 presents the (partial) pitch contours that result from adding, for nine words, the partial effect of WORD to the base contour for females. (Contours for male speakers are similar, and are thus not shown.) For most words, the fall-rise-fall pattern is preserved. However, there are substantial differences in how F0 varies with time. While some words have a pronounced initial fall, e.g., (c) and (d), this fall can be severely muted, e.g., (f) and (i). Even though all these words are supposed to be realized with a fall-rise-fall pattern theoretically derived from an underlying RF tone specification, the details of actual phonetic realizations differ consistently on a word-by-word basis.

3. ENGLISH

In English, pitch contributes to the perception of lexical stress and plays an important role at the

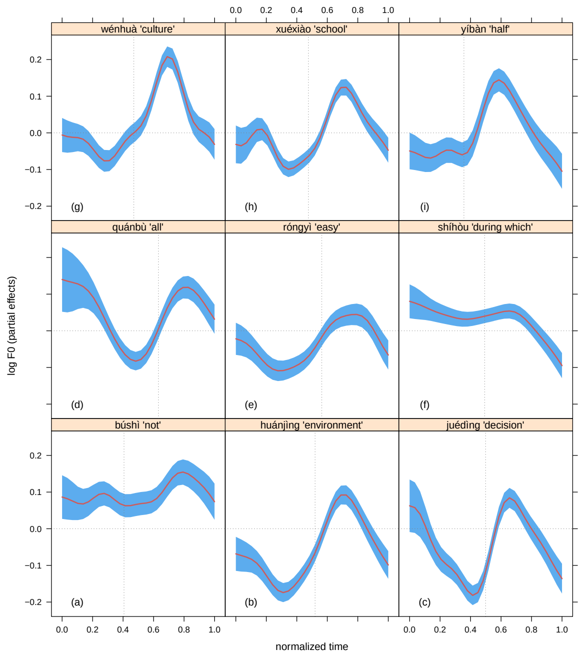

Figure 2: Predicted (partial) contours for a sample of nine Mandarin words. Vertical dotted lines indicate average syllable boundaries.

discourse level, for example to distinguish between questions and statements or between old and new information. It has been suggested that English has between four and six types of meaningful pitch accents [10, 11, 12], largely identified by the alignment of F0 peaks and valleys with segmental material [13]. However more recent research has shown that the alignment of peaks shows considerable variation between speakers, for the same speaker on different occasions and with the segmental composition of words [14, 15]. The relationship between pitch contour and function is therefore probabilistic rather than categorical [16].

3.1. Data and analyses

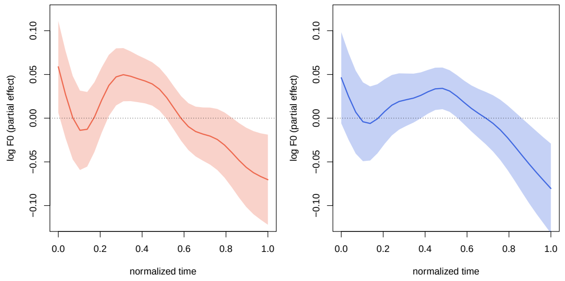

Figure 3: Predicted partial contours of English left-stressed disyllabic words for females (left) and males (right).

The English data were obtained from the Buckeye Corpus [17]. We selected the 72 left-stressed disyllabic words that have at least 20 occurrences in this corpus, producing a total of 4,724 tokens. As with Mandarin, F0 measurements were taken across the entire pitch contour, and measurement timepoints were normalized by duration. Local speech rate and utterance position were also calculated. Instead of the tonal context variable used for Mandarin, we coded, for each token, the stress level of the immediately preceding and following syllables (no stress, primary stress, or secondary stress, according to the CMU Pronouncing Dictionary[3]). We fitted a GAMM to the English F0 data, using the same model specification that was used for the Mandarin data.

3.2. Results

The partial pitch contours predicted for the English words are shown in Figure 3, differentiated by `SEX`. As with the Mandarin data, the female and male contours are similar, with the contours for females varying slightly more in magnitude than those for males. However, compared to Mandarin, the predicted contours come with wider confidence intervals, indicating more variability in the realization of pitch. Although the contours feature a small dip at the beginning followed by a gradual fall, only the falling part should be considered, given that the confidence interval of the initial fall-rise always contains 0. The pitch contour of left-stressed words therefore emerges as initially basically level, followed by a small rise and then a long shallow fall. Inclusion of a by-word factor smooth resulted in a substantial increase in goodness of fit (by 36,180 AIC units). We carried out the same randomization experiment as for Mandarin, obtaining a distribution with $M = 13,504$ and $SD = 1,292$, confirming that the random effect of `WORD` is unlikely to be a statistical artefact.

Figure 4 presents the predicted (partial) contours for nine English words. Compared with the pitch realizations of Mandarin words, the contours for the English words appear to be more idiosyncratic. For some words, e.g., (c) and (e), a falling pitch is obviously present, but for others, e.g., (d) and (f), some kind of plateau can be observed. The contour of the word ‘only’ (h) is to a large extent a flat line, while the contour for ‘business’ (b), has two rises.

4. GENERAL DISCUSSION

In this study we investigated pitch variation in Mandarin and English words. For both languages, word identity is a strong predictor, accounting for

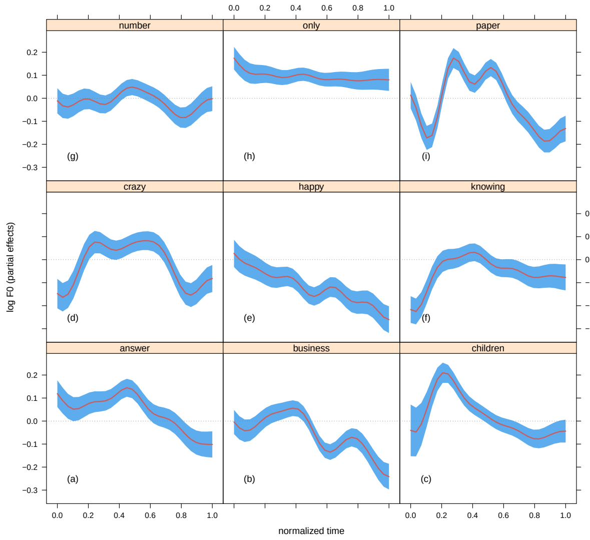

Figure 4: Predicted partial contours for a sample of nine English words. The contours were obtained by summing the base contour for females and the word-specific partial contours.

a significant amount of variance in F0, even for words with theoretically the same underlying pitch contour. This result suggests that there is by-word systematicity in pitch realization, and that in both languages, words appear to have their own pitch signatures.

What differs between the two languages is that in Mandarin, the signatures are ‘variations on a theme’ provided by a sine-wave like tone contour, whereas in English, word-specific signatures are more idiosyncratic, and show less resemblance to the general pitch contour, at least for disyllabic left-stressed words. This cross-linguistic difference may be due to the fact that compared to Mandarin tones, stress in English is more loosely tied to pitch, and that English words’ actual pitch realizations are largely dependent on utterance intonation. However, in further analyses with additional discourse-based variables, we have so far not found a single predictor or a combination of predictors that can completely replace the by-word effect.

Despite our results, it might be thought that the evidence for word-specific pitch contours will evaporate once all variables known to co-determine pitch have been taken into account. But this argument runs into at least three problems. Firstly, such a hypothesis would be difficult to test, since many of the factors co-determining pitch are highly correlated, leading to substantial collinearity and concomitant lack of interpretational transparency. Secondly, the argument is logically self-defeating since the very large number of factors known to influence pitch effectively define the word type and possibly even the specific token. Thirdly, this line of reasoning supposes a straightforward causal link between the full array of words’ other properties (lexical, syntactical, and discourse-related, but excluding pitch) and the observed word-specific pitch signatures. However, such a unidirectional causal explanation seems unlikely, since pitch is an integral part of every token heard.

Especially in the light of recent advances in artificial intelligence that use statistical patterns across the speech signal as a whole, it makes sense to consider the present findings in a broader perspective, where the word as a fundamental unit plays a much more modest role. We note that words’ forms are subject to immense variation, not only at the level of segments and syllables, for which reduction is widespread [18], but also at subsegmental levels [19]. Furthermore, we note that what a word actually means is also highly dependent on the context in which it is used [20, 21, 22]. It follows that the word as a theoretical and cultural construct suggests far more constancy and unity than is actually present in spoken language. This conclusion is supported by studies reporting a much tighter link between fine-grained phonetic detail and shades of meaning than is generally assumed to exist [23, 24, 25].

The finding of word-specific pitch contours therefore does not entail or suggest that a given word is always realized in exactly the same way. Just as there is variation in the realization of a Mandarin tone type or an English stress pattern, we assume there will be variation in the exact realization of word types, reflecting large numbers of parameters including e.g., lexis, syntax, discourse, pragmatics, and emotion that jointly determine not only pitch, but also segmental realization, assimilation, and co-articulation. What the GAMM provides is a best estimate of the pitch contour of a word averaged across all its usages. We therefore hypothesize that the word-specific pitch signatures revealed by our models are the pitch contours with which words are most likely to be realized. They might be seen as reflections of, or pointers to, words’ most typical patterns of use.

In conclusion, we note that from a learning perspective word-specific pitch contours make words more discriminable from one another and might hence facilitate learning. At a practical level, they could therefore have useful application in language teaching, where it may be helpful to draw attention to words’ most characteristic contours.

5. REFERENCES

[1] A. T. Ho, "The acoustic variation of Mandarin tones," *Phonetica*, vol. 33, no. 5, pp. 353–367, 1976.

[2] I. Plag, G. Kunter, and M. Schramm, "Acoustic correlates of primary and secondary stress in north american english," *Journal of Phonetics*, vol. 39, no. 3, pp. 362–374, 2011.

[3] S. N. Wood, *Generalized Additive Models*. New York: Chapman & Hall/CRC, 2017.

[4] Y. Xu, "Sources of tonal variations in connected speech," *Journal of Chinese Linguistics Monograph Series*, pp. 1–31, 2001.

[5] C. Shih, "Tone and intonation in mandarin," *Working Papers, Cornell Phonetics Laboratory*, vol. 3, pp. 83–109, 1988.

[6] Y. Xu, "Contextual tonal variations in Mandarin," *Journal of Phonetics*, vol. 25, no. 1, pp. 61–83, 1997.

[7] C. Cheng and Y. Xu, "Mechanism of disyllabic tonal reduction in Taiwan Mandarin," *Language and speech*, vol. 58, no. 3, pp. 281–314, 2015.

[8] J. Fon, "A preliminary construction of Taiwan Southern Min spontaneous speech corpus," National Science Council, Taipei, Taiwan, Tech. Rep. NSC-92-2411-H-003-050-, 2004.

[9] F. Tomaschek, P. Hendrix, and R. H. Baayen, "Strategies for addressing collinearity in multivariate linguistic data," *Journal of Phonetics*, vol. 71, pp. 249–267, 2018.

[10] M. E. Beckman and G. Ayers, "Guidelines for ToBI labelling," 1997, http://www.cs.columbia.edu/~agus/tobi/labelling_guide_v3.pdf.

[11] E. Grabe, B. Post, and F. Nolan, "Modeling intonational variation in english," *The IViE system. Paper presented to the Proceedings of Prosody*, 2000.

[12] L. C. Dilley and C. C. Heffner, "The role of f0 alignment in distinguishing intonation categories: evidence from american english," *Journal of Speech Sciences*, vol. 3, no. 1, pp. 3–67, 2013.

[13] J. Barnes, N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel, "Tonal center of gravity: A global approach to tonal implementation in a level-based intonational phonology," *Laboratory Phonology*, vol. 3, no. 2, pp. 337–383, 2012.

[14] S. Calhoun, "The centrality of metrical structure in signaling information structure: A probabilistic perspective," *Language*, pp. 1–42, 2010.

[15] J. Cole and S. Shattuck-Hufnagel, "New methods for prosodic transcription: Capturing variability as a source of information," *Laboratory Phonology*, vol. 7, no. 1, 2016.

[16] C. Kurumada and T. B. Roettger, "Thinking probabilistically in the study of intonational speech prosody," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 13, no. 1, p. e1579, 2022.

[17] M. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, "Buckeye corpus of conversational speech (2nd release)," 2007, columbus, OH: Department of Psychology, Ohio State University (Distributor). [Online]. Available: www.buckeyecorpus.osu.edu

[18] K. Johnson, "Massive reduction in conversational American English," in *Spontaneous speech: data and analysis. Proceedings of the 1st session of the 10th international symposium*. Tokyo, Japan: The National International Institute for Japanese Language, 2004, pp. 29–54.

[19] Y. Yao, "Understanding VOT variation in spontaneous speech," *UC Berkeley PhonLab Annual Report*, vol. 5, no. 5, 2009.

[20] J. L. Elman, "On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon," *Cognitive science*, vol. 33, no. 4, pp. 547–582, 2009.

[21] J. R. Firth, *Selected papers of J R Firth, 1952-59*. Indiana University Press, 1968.

[22] T. Landauer and S. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge," *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.

[23] S. Gahl, "Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech," *Language*, vol. 84, no. 3, pp. 474–496, 2008.

[24] K. K. Drager, "Sociophonetic variation and the lemma," *Journal of Phonetics*, vol. 39, no. 4, pp. 694–707, 2011.

[25] S. Gahl and H. Baayen, "Time and thyme again: Connecting spoken word duration to models of the mental lexicon," *arXiv*, 2022. [Online]. Available: https://osf.io/2bd3r/

[26] S. N. Wood, "Thin plate regression splines," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 1, pp. 95–114, 2003.

[1] This research was funded by ERC, project SUBLIMINAL (No. 101054902).

[2] TPRS refers to thin plate regression splines [26].

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict