Language (2026), **102**:1, 1–45
doi:10.1017/S0097850725000001

Linguistic Society of America

GENERAL RESEARCH ARTICLE

Word-specific tonal realizations in Mandarin

Yu-Ying Chuang[1], Melanie J. Bell[2], Yu-Hsiang Tseng[3], and R. Harald Baayen[3]

[1]Department of Taiwan Culture, Languages and Literature, National Taiwan Normal University, Taiwan
[2]Anglia Ruskin University, Cambridge, UK
[3]Department of Linguistics, University of Tübingen, Tübingen, Germany

Corresponding author: Yu-Ying Chuang; Email: yuying.chuang@ntnu.edu.tw

Received: 04 May 2024; **Revised:** 10 October 2025; **Accepted:** 19 November 2025

Keywords: tone; Mandarin; embeddings; GAMs; form-meaning isomorphy; rise-fall tone pattern; two-syllable words

Abstract

The pitch contours of Mandarin two-character words are generally understood as being shaped by lexical tones on the constituent single-character words, in interaction with articulatory constraints imposed by factors such as speech rate, coarticulation with adjacent tones, segmental makeup, and predictability. This study shows that tonal realization is also partially determined by words' meanings. We first show, on the basis of a corpus of Taiwan Mandarin spontaneous conversations, using a generalized additive regression model and focusing on the rise-fall tonal pattern, that after controlling for effects of speaker and context, word type is a stronger predictor of tonal realization than all of the previously established word-form-related predictors combined. Importantly, the addition of information about meaning in context improves prediction accuracy even further. We then proceed to show, using computational modeling with context-specific word embeddings, that token-specific pitch contours predict word type with 50% accuracy on held-out data, and that context-sensitive, token-specific embeddings can predict the shape of pitch contours with 40% accuracy. These accuracies, which are an order of magnitude above chance level, suggest that the relation between words' pitch contours and their meanings are sufficiently strong to be potentially functional for language users. The theoretical implications of these empirical findings are discussed.

1. Introduction

The tone system of Mandarin Chinese is commonly described as having four lexical tones, namely high level (T1), rising (T2), dipping/low (T3), and falling (T4), plus one neutral or floating tone, whose shape depends on the preceding tone (Chao 1968). While these tonal representations are well established in the relevant literature and are taught both in Chinese schools and in second language learning classrooms, it is also known that their realizations, that is, the actual pitch contours of spoken words, can differ significantly from the canonical descriptions.

In fact, the mismatch between words' actual realizations and their canonical forms is ubiquitous. Substantial variability can be observed when one investigates how a word is pronounced in spontaneous speech. Johnson (2004), for example, reports that in English conversational speech, the word *until* is pronounced as [ʌntɪl], [ʌntəl], [ɛntɪl], [ɛntəl], [ɪntɪw], [n̩tɪl], [ənt̬ʌ], [n̩t̬l̩], [t̬l̩], and [tə]. That is, a word represented by a single orthographic form can undergo massive syllable deletion and segment deviation, leading to multiple pronunciation variants. The same holds for Mandarin. Chung (2006) observed that the

© The Author(s), 2026. Published by Cambridge University Press on behalf of the Linguistic Society of America. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (http://creativecommons.org/licenses/by-nc-sa/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

connective *sui1ran2* [sʷeɪzan] 'although' is often pronounced as [sʷeɹa] or [sʷeɪm], and *ke3shi4* [kʰəʂɨ] 'but' as [kʰəɻ] or simply [kʰəː].[1] About one-third of the syllables produced in spontaneous Mandarin are contracted syllables; that is, they are reduced pronunciations of multisyllabic words (Tseng 2005b). It can thus be seen that the phonetic realizations of a word vary, and the canonical description of a word's phonological form is rarely encountered in conversational speech.

Not only do tokens of the same word vary substantially in the realization of their segments, but, in tonal languages, they also vary substantially in the realization of their lexical tone. Xu (2001) distinguishes two sources of tonal variation: voluntary and involuntary. Voluntary variation arises from the speaker's communicative intentions, such as eliciting a question or creating emphasis, where changes in rhythm, accent placement, and intonation can all give rise to substantial modification of tonal realization; these voluntary constraints are intimately related to the syntactic and pragmatic functions of an utterance (Gårding 1987, Shen 1989, 1990a, Xu 1999, Liu & Xu 2005). In addition, the realization of tone is modulated by sociolinguistic and paralinguistic factors such as gender, dialect, and emotion (Fon & Chiang 1999, Zhang et al. 2006).

Involuntary variation in tone production arises from articulatory constraints posited to be beyond the speaker's control. Some of these constraints are related to the processes of connected speech. At the utterance level, where a sequence of tones is produced, the realization of a given tone is greatly influenced by its preceding and following tones, leading to tonal coarticulation (Shih 1988, Shen 1990b, Xu 1997). Coarticulated tones usually deviate from their canonical tonal shapes; in extreme cases the original shapes are no longer preserved (Shen 1989, Shih & Kochanski 2000) and may be unrecognizable to native speakers (Xu 1994). Physically speaking, it takes a certain amount of time to raise or lower pitch (Xu & Sun 2002), and therefore tonal realizations are highly dependent on whether speakers have sufficient time to realize a given tonal contour. Under the time pressure of fast speech, tonal targets are often not fully realized (Tang & Li 2020); this leads to significant deviation from the canonical patterns and often results in tonal reduction (Cheng & Xu 2015). The accepted descriptions of the tones as 'level', 'rising', 'dipping/low', and 'falling' are therefore generalizations across considerable variability in the fine detail of their phonetic realizations.

In addition to the effects of connected speech, articulatory constraints on tone production arise from the segmental makeup of syllables. At the syllable level, vowels, onsets, and syllable structure are all known to contribute to tonal variation (Howie 1974, Ho 1976, Whalen & Levitt 1995, Xu & Xu 2003, Fon & Hsu 2007). Over and above contextual effects, it is therefore clear that different words with the same canonical tone will differ in the details of their tonal contours, since they differ in their segmental makeup. In previous studies on Mandarin lexical tones using laboratory speech, by-word variation has rarely been taken into account, since very often the same words are used across different experimental conditions for maximum control (e.g. Chen 2010, Li & Chen 2016). Previous studies using corpus data have accommodated by-word variation as a random effect in the relevant statistical models. For example, Wu et al. (2020) worked with random intercepts for word form. However, this treats the word as a source of noise, where different words exhibit idiosyncrasies that are irrelevant to the predictors of interest. In contrast, other studies have specifically investigated how word-level properties affect tone production. For example, Zhao and Jurafsky (2009) showed that usage frequency affects f0 realization in Cantonese words with mid-level or mid-rising tones.[2] For Mandarin, Tang and Shaw (2021) showed that f0 is affected not only by word frequency but also by informativity, defined as a word-level variable. To the best of our knowledge, however, no previous study has specifically addressed the relationship between tonal realization and word meaning. This article fills that gap.

In the extensive literature on Mandarin tones, their semantic function is assumed to be straightforward: different tones distinguish between alternative meanings. For instance, *jiu4* 'then' and *jiu3* 'nine' are differentiated by having a falling and a dipping tone, respectively. However, the very same combinations

[1] Throughout this article, Pinyin transcriptions are followed by numeric notations representing the four tones.

[2] Zhao and Jurafsky (2009) describe Cantonese as having six lexical tones.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

of segments and tone often realize many other different meanings, as exemplified by *jiu3* 'nine' and *jiu3* 'alcoholic beverage'. The combination of strong phonotactic constraints on syllable structure and a limited number of lexical tones has given rise to widespread homophony and polysemy, often in combination with homography. For instance, *jiu4* 'then' has a wide range of translation equivalents in English, including *then*, *at once*, *only*, *already*, *to approach*, *to accomplish*, *to suffer*, and *to take advantage of* (https://www.pleco.com/, s.v.).

The examples in the preceding paragraph are all monosyllabic words; however, in the Chinese Lexical Database (Sun et al. 2018), only 8% of the 48,000 words are monosyllabic. The majority (70.4%) of Mandarin words are disyllabic, written with two characters.[3] The tonal targets of disyllabic words are taken to be subject to the same voluntary and involuntary constraints that govern the realization of tone in monosyllabic words. As a consequence, all disyllabic words sharing, for example, an initial falling tone and a subsequent rising tone are assumed to have the same underlying pitch contour; any differences in how the tones are realized are assumed to be attributable to the involuntary and voluntary processes described above. However, the present study will show that, alongside the known articulatory constraints, there is a previously undocumented close association between the meanings of Mandarin disyllabic words and the realization of their tonal contours.

The basis for our study was laid by a growing body of research on English showing that fine-grained phonetic variation can be systematically associated with differences in meaning. For example, at the word level, Gahl (2008) showed that homophones such as *time* and *thyme* are realized on average with different acoustic durations. In the same vein, Lohmann (2018) found that the durations of words such as *cut* depend on whether they are used as nouns or verbs. These differences in word duration were initially explained as a consequence of the different relative frequencies of the homophones in these studies. However, Gahl and Baayen (2024) showed that the meanings of English homophones are a strong co-determinant of their spoken word durations even after frequency differences and other co-determinants such as speech rate are taken into account. A relationship between meaning and duration has also been found for the English suffix /s/: different grammatical functions of this suffix (e.g. plural and third-person singular) tend to be realized with different durations (Plag et al. 2017). Furthermore, the relationship between meaning and phonetic realization may extend beyond durational differences. Drager (2011) reported that the phonetic realization of the word *like* varies according to its discourse or grammatical meanings, not only in the duration of the consonants but also in the degree of diphthongization of the vowel.

The results described in the previous paragraph are compatible with a theory of the mental lexicon that postulates a direct connection between the context-specific meaning of a word token and the details of its phonetic realization. Such a theory has been computationally implemented as the **Discriminative Lexicon Model**, henceforth DLM (Baayen et al. 2019, Chuang & Baayen 2021, Heitmeier et al. 2026). In this model, lexis and morphology are acquired through a process of error-driven learning that allows for fine-grained alignments between low-level properties of form and low-level properties of meaning, both operationalized as high-dimensional numeric vectors. The model captures relationships between these vectors in two networks: a comprehension network that maps word form onto word meaning, and a production network that maps word meaning onto word form. In other words, given a word-form vector, the comprehension network is able to compute a predicted meaning vector, analogous to the comprehension process in which we recognize and make sense of the meaning of a word according to the visual or auditory input. In the same vein, the production network computes a predicted form vector given a semantic vector, analogous to the production process where we produce a form to express an intended word meaning. It has been shown that, across a range of languages, these networks successfully capture an alignment between meaning and fine-grained variation in form, for example, word duration in

[3] In this study we use a corpus of Taiwan Mandarin spontaneous speech (see Fon 2004) and take the word labels supplied by the corpus as given. The labeled words include, for example, nouns such as *xue2xiao4* 'school', verb forms such as *xue2dao4* 'learn +resultative', and negated verbs such as *bu2shi4* 'not+be'. Since all of these forms have two syllables, we refer to them collectively as disyllabic words.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

Mandarin (Chuang et al. 2023), different durations of homophones in English (Gahl & Baayen 2024), and the degree of tongue lowering in the articulation of the [a] vowel in German (Saito et al. 2023).

Despite being able to predict both word forms and word meanings, the DLM model does not store whole-word representations of either kind; rather, the model's memory consists of the connection weights in the networks, which are continuously recalibrated with each learning event (see Heitmeier et al. 2023 for experimental evidence). Likewise, in the corresponding theory of the mental lexicon, word forms and meanings do not have representations in memory. The forms are ephemeral auditory or visual experiences, which dynamically generate corresponding, equally ephemeral, meaning representations. Conversely, a meaning conceptualized by a speaker at a given point in time is dynamically transformed into ephemeral representations driving articulation. In other words, the DLM posits a lexicon in which lexical items are neither static nor discrete. Rather, the lexicon is taken to consist of a series of dynamic, modality-specific neural networks (Baayen et al. 2019) that are constantly fine-tuned, by adjusting connection weights, in order to optimize word comprehension and production. For further technical details of the implementation of the DLM mappings, please refer to the supplementary material.[4]

At this point, the question arises of how to understand the linguistic term 'word'. In this study, we define **word token** as a pairing of a specific form with a specific meaning. We define **word type** (or simply **word**) as a set of word tokens that have the property of having both similar forms and similar meanings. For instance, the set of phonetic realizations of *jiu3* and their corresponding context-specific meanings ('wine, liquor, spirits, alcoholic beverage'; https://www.pleco.com/, s.v.) jointly constitute the tokens of the word type *jiu3*.[5] This working definition of 'word' seeks to do justice to the fact that no two tokens of the same word, as produced by humans, are ever completely identical in form. It also adopts the theoretical position of distributional semantics, namely, that what a word means varies with its context (Harris 1954, Firth 1968, Landauer & Dumais 1997, Elman 2009). Thus, in the framework of the DLM, words are sets of input-output pairs on which the production and comprehension networks are trained, which leave 'traces' in the connection weights of these networks, but which are themselves not stored as independent entities.

As will be apparent from the preceding paragraph, our theoretical position apropos of meaning is one of contextualism. We assume that utterances rather than sentences are the domain of propositional content and that the meaning of an utterance depends on the context in which it is produced. Our notion of context includes not only the content, genre, and style of the surrounding text but also extralinguistic factors such as who is speaking and who they are addressing. Hence, our view of meaning does not draw a neat boundary between semantics and pragmatics.

Both the theoretical assumptions of the DLM and the empirical studies of English homophonous words and affixes by Gahl and Baayen (2024) and Plag et al. (2017), respectively, suggest the possibility that Mandarin homophones also differ systematically in phonetic detail, that is, that their segments and/or tones are realized slightly differently according to the intended meaning. This could apply to homographic pairs such as *da4jia1* 'everyone' and 'art master', as well as to nonhomographic pairs such as *shu4mu4* 'tree' and 'number'. In other words, it is possible that the realizations of canonical tones are determined not only by the involuntary and voluntary constraints previously described, but also by the context-specific meanings of the word tokens on which they are realized. If this is correct, then conversely it is not only the four canonical pitch contours that help to distinguish between alternative meanings, but also the finer details of their phonetic realization. This brings us to the central hypothesis of our study.

(1) Hypothesis: The unique pitch contour of each spoken Mandarin word token is determined in part by the specific meaning of that token. These meaning-specific contours are in principle learnable and hence potentially used by speakers.

[4] The supplementary material, including further documentation of specific issues and also the code for statistical analyses and computational modeling in this article, is available at https://osf.io/nwv74/.

[5] Note that, in this case, the Chinese orthographic character can be used to represent the word, since the character of *jiu3* is not homographic: all of its meanings cluster around the concept of 'alcoholic beverage'.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

In what follows, we provide evidence for four more specific predictions that we derive from this hypothesis.

(2) Predictions
   (i) Word type will be a stronger predictor of tonal realization than all of the previously established word-form-related predictors combined. This prediction follows from the hypothesis because word type includes information about meaning in addition to information about form.
   (ii) Information about a word's meaning in context, that is, its sense, will improve prediction of its tonal realization, compared with prediction based on word type alone. This prediction follows from the hypothesis because a word's senses include all of the information encoded in the word, plus additional more fine-grained semantic information. It therefore allows us to tease apart effects of meaning from possible effects of form.
   (iii) Given a pitch contour, the meaning of its carrier token can be predicted above chance level by a simple computational model with previous experience of that word type. This prediction follows from the hypothesis that meaning-specific contours are in principle learnable.
   (iv) Assuming it has previous experience of the relevant word type, a simple computational model can produce an appropriate pitch contour for a given meaning. This also follows from the hypothesis that meaning-specific contours are in principle learnable.

In this article we explore these predictions for disyllabic words with the canonical tone specification of a rising tone (T2) followed by a falling tone (T4), henceforth the rise-fall tonal pattern, RF. The disyllabic word is a natural choice for our study, since Mandarin vocabulary is composed mostly of disyllabic words (Huang et al. 2010, Wu et al. 2023). We decided to focus on the RF pattern because it is the heterogeneous tonal combination with the highest number of word types and tokens in the speech corpus we used. We wanted to investigate a heterogeneous tonal combination rather than a homogeneous one to ensure that the results obtained are not specific to a given tone.

The remainder of the article proceeds as follows. Section 2 addresses the first two predictions listed above. It describes how we used generalized additive modeling to analyze the pitch contours of RF words extracted from the above-mentioned corpus of spontaneous speech. We discuss our modeling strategy and present the results of an analysis based on word type and one enhanced with word sense. Section 3 addresses the third and fourth predictions. It describes how we used computational modeling with the DLM and distributional semantics to demonstrate that meaning-specific pitch contours have the potential to facilitate comprehension and to be produced in response to intended meaning. We discuss the implications of our results in Section 4.

2. Establishing word- and meaning-specific pitch contours

This section describes how we addressed the first two predictions outlined in Section 1. We modeled the pitch contours of spoken tokens of Mandarin disyllabic words with the RF tonal pattern, using generalized additive modeling. To explore prediction (i), we evaluated the effectiveness of word type as a predictor of tonal realization, compared with the segment-related articulatory constraints previously described in the literature. To explore prediction (ii), we evaluated whether adding information about a word's meaning in context would improve prediction of tonal realization, compared with prediction based on word type alone. Sections 2.1 to 2.4 describe aspects of the methodology, Sections 2.5 to 2.7 report the results of the generalized additive models, and Section 2.8 summarizes Section 2 overall.

2.1. Generalized additive modeling

Classical analyses of pitch typically take measurements at various contour landmarks, such as maximum and minimum f0 values. However, since pitch actually varies continuously with time, such analyses miss

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

much of the detail of the f0 contour. To better capture the complete shapes of tonal variations, we modeled the pitch contours of the tokens in our data using generalized additive models, henceforth GAMs (Wood 2017). GAMs relax the regression assumption that the relation between a predictor and response should be linear; instead, the model incorporates individual, potentially nonlinear relationships between each predictor variable and the response variable. For main effects, this relationship is estimated using functions known as smoothing splines (henceforth smooths), which can fit either a line or a (possibly wiggly) curve to the data, as required. Nonlinear interactions can be included using functions called **tensor product smooths**, which fit a wiggly (hyper)surface for the joint effect of two or more predictors. In addition, it is possible to include nonlinear random effects, for instance by using functions called **factor smooths** (Baayen et al. 2022), which fit a wiggly curve for each level of the random factor, for example, for each individual speaker in the case of a by-speaker factor smooth. Because GAMs model complex nonlinear relationships, they make it possible to model f0 as a nonlinear function of time across an utterance, while also including other predictors known to affect pitch, such as speech rate and speaker gender. They can thus capture fine-grained modulations of pitch as time unfolds.

To illustrate our GAM-based modeling strategy, we created a toy data set consisting of six disyllabic Mandarin word types with the RF tonal pattern. Their audio files were downloaded from Meng Dian, a publicly available Taiwanese online dictionary. In the audio files, a single female speaker pronounces each word twice, so that we had a total of twelve tokens, all from the same speaker. We used the **To Pitch** (CC) command in Praat (Boersma & Weenink 2019) to estimate f0 values for each token. Because the speaker is female, we set the pitch floor at 50 Hz and the pitch ceiling at 400 Hz. To optimize the f0 estimation, we set the time step at 0.001 seconds and used the most accurate method available, namely a Gaussian window; other parameters were left at the default values. This gave us f0 values at one millisecond intervals for the voiced sections of the tokens. Next, we used Praat's **To PointProcess** command to obtain the time points of the glottal pulses in the voiced sections of the tokens. We then extracted the f0 values corresponding to the time points of the glottal pulses. Since the time points of glottal pulses do not necessarily correspond exactly to the one millisecond measurement intervals, we used linear interpolation between adjacent measurements to estimate f0 for the glottal pulses. At this stage no f0 values were interpolated for the voiceless sections of the tokens. Because the tokens varied in duration, we transformed the time points of the measurements onto a normalized time scale of 0–1.[6] The f0 values of one of the two renditions of each of the six words are plotted on this normalized time scale in the left-hand panel of Figure 1. Although there are no data points for the voiceless segments, it can be seen that the RF tonal sequence in Mandarin is realized with a small initial fall, followed by a rise, and finally a much larger fall. The dipping realization of T2 is consistent with previous findings for laboratory speech that the rising portion of Mandarin T2 is usually preceded by a slight fall (Ho 1976, Tseng 1981, Shih 1988, Shen & Lin 1991, Moore & Jongman 1997, Xu 1997).[7]

Using the mgcv package (Wood 2017) for R (R Core Team 2022), we fitted a GAM to the toy data set, with f0 as the dependent variable and normalized time as the only predictor. Including time as a predictor allows us to model the entire pitch contour of a tonal pattern by predicting f0 across the whole timespan of a token of that pattern. The GAM predicts f0 values not only for voiced sections but also for voiceless parts of the tokens by estimating what the pitch contour would likely be if the voiceless segments were voiced. In adopting this modeling strategy, our goal is not curve fitting but cognitive modeling of pitch contours (see also prosody models such as Kochanski & Shih 2003, Fujisaki 2004, and Prom-On et al. 2009). We theorize that a speaker producing a word has a pitch contour for the whole word or even the whole

[6] Time normalization enables us to focus on the shape of pitch contours. In order to take into account possible effects of word duration on the realization of tone, in the more comprehensive analyses reported below, token duration is included as a covariate, in interaction with normalized time. Note that by reversing the time normalization, the predicted curves can straightforwardly be back-transformed to the original time scale.

[7] The initial fall could also reflect dialectal variation specific to Taiwan Mandarin, since in this language, T2 is predominantly realized with a concave contour. This concave contour may have become a standardized realization that no longer reflects articulatory constraints (see e.g. Fon & Hsu 2007).

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

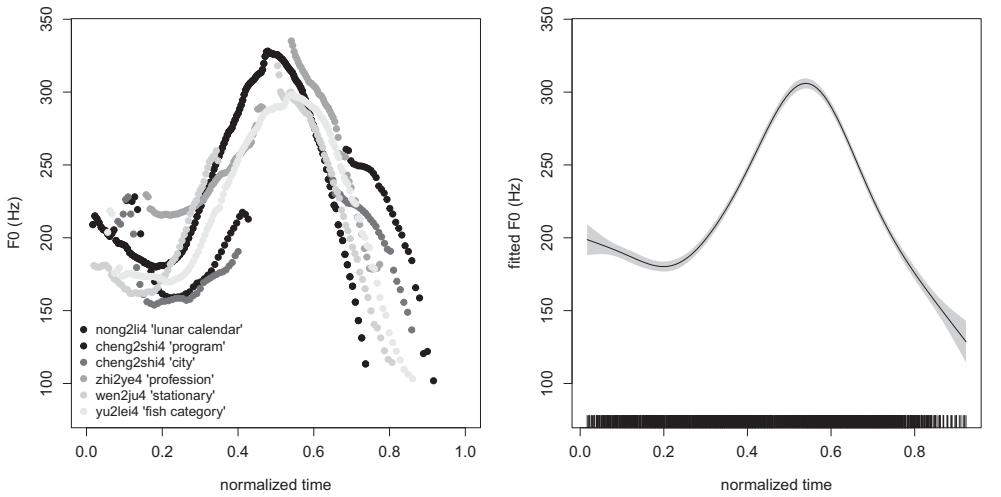


Figure 1. *Toy data set. The left-hand panel shows the f0 contours of single tokens of six Taiwan Mandarin words with the RF tonal pattern, produced in isolation by the same speaker. The right-hand panel shows the RF contour predicted by a simple GAM, using a thin plate regression spline smooth for normalized time as predictor.*

utterance in mind. We further assume that the cognitive planning underlying the production of pitch contours is continuous, rather than a step function with jumps to 0 Hz for voiceless segments. From our theoretical perspective, it is only at the stage of articulation that voiceless segments mask the internally projected pitch development.

The pitch contour predicted by the GAM, shown in the right-hand panel of Figure 1, captures the general trend with some precision, mirroring the raw data on the left. This graph is the model's best estimate of the average population contour for words with the RF tonal pattern, given the twelve tokens in our toy data set. However, the empirical contours show considerable variation around this average, even for a single speaker producing citation forms in isolation. This variability in realization is the focus of our study.

To investigate whether individual tonal realizations might be related to the meaning of the carrier word token, we enriched the model with a by-word factor smooth, which is effectively a nonlinear, time-dependent, random effect for word type. In the present example, provided purely for illustration of the method, the by-word smooths are based on just two tokens of each word.[8] This mixed model predicts, for each word type, a word-specific adjustment contour that has to be added to the population pitch contour to obtain the predicted pitch for a given word type.

The by-word adjustment contours estimated by the GAM are visualized in the left-hand panel of Figure 2. The dotted line at $y = 0$ is a reference line: an adjustment curve for a given word that followed this line would indicate that no adjustment is needed and that this word's pitch is identical to the population contour. Deviations above this reference line indicate an upward f0 adjustment, and deviations below it indicate a downward adjustment. The word *zhi2ye4* 'profession', for example, represented by a dashed curve, requires an upward adjustment for the entire contour, although the amount of adjustment varies across time. When a given word's adjustment contour is added to the general contour, we obtain its fitted contour, as shown in the right-hand panel of Figure 2. The dashed line in each graph represents the general contour, which is, by definition, the same for all words. The black line (along with its confidence interval in gray) plots the fitted contour for the word in question. These fitted contours vary from word to word. For

[8] For detailed discussion of the ways in which these smooths can be specified, and the accuracy of these smooths, see Baayen et al. 2022.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

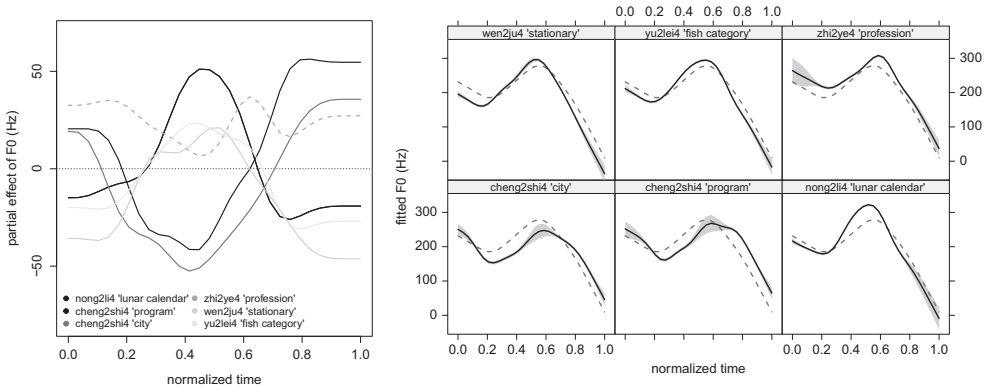


Figure 2. *The left-hand panel shows by-word adjustment contours from the toy model with only by-word factor smooth and normalized time as predictors. The right-hand panel plots the fitted contour for each word, with the predicted general contour (identical for all words) indicated by the dashed line.*

zhi2ye4 'profession', presented in the right-most upper panel, the entire fitted contour is above the general average contour, as expected. The homophone pair *cheng2shi4* 'city' and *cheng2shi4* 'computer program', shown in the left and middle lower panels, have similar but not identical fitted contours, as would be expected if word meaning, as well as word form, plays a role in determining tonal realization.

Because the GAM receives no input for voiceless sections of the tokens, where there is no actual pitch, the model's estimations for these sections are more uncertain than those for the voiced sections. Similarly, word-specific partial effects for voiceless sections tend toward 0, that is, no effect, again because there is no data. For instance, in Figure 2, the panel for *cheng2shi4* 'program' shows that the GAM produces a wide confidence interval for the two voiceless consonants, exactly as it should. Because there is no data for the voiceless segments of a given token, the model falls back on the overall smooth of time in these intervals, that is, the general rise-fall tonal pattern. Hence, in Figure 2, it can be seen that the wide confidence intervals for voiceless sections of the tokens tend toward including the dotted line that represents the average contour.

Although this is just a toy example, it does illustrate two important aspects of the more detailed analyses reported below. First, we can decompose the observed pitch contour of any token into a general population contour plus various more specific contours, including a meaning-specific adjustment contour. Second, GAMs can identify such meaning-specific contours and, given an adequate sample size, they could inform us about whether including meaning-specific contours improves model fit.

In what follows, we turn to a much larger data set of spontaneous spoken Taiwan Mandarin and consider a much broader set of predictors that allow us to bring under statistical control a wide range of constraints known to co-determine the realization of pitch. If there is indeed a semantic component to the tonal realization of Mandarin disyllabic word tokens, then by-word factor smooths should be well supported even when relevant control variables are taken into account.

2.2. Data

We used the Taiwan Mandarin spontaneous speech corpus, which is one of a pair of corpora created by researchers at the National Taiwan University (see Fon 2004 for a description of the general method in the context of the Southern Min corpus). The Mandarin corpus consists of about thirty hours of recorded interviews with fifty-five native Taiwan Mandarin speakers aged between twenty and sixty years, thirty-one self-identified as female and twenty-four self-identified as male, recruited through snowball sampling starting with friends, acquaintances, or relatives of the team. The recordings were made over several years between 2000 and 2010, using high-quality microphones and digital recorders in quiet locations,

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

whenever possible in a soundproof laboratory. Before the interview, participants were told that the purpose was to understand their views on changes in various aspects of life. The interviewers aimed not to engage in conversation with participants but to elicit longer monologues about their personal experiences in childhood, school, work, relationships, and elsewhere. After the interview, participants were debriefed and asked for permission to include their recording in the corpus. If permission was granted, the researchers orthographically transcribed the recording using Chinese characters. The word segmentation system developed by Academia Sinica (Ma & Chen 2003) was used to identify word boundaries in the orthographic representation, and dictionaries were used to add information about the canonical tones of the words identified. The character transcriptions were then romanized in order to run a forced aligner (Easyalign: Goldman 2011) to match the transcription to syllable boundaries in the audio files. The resulting alignment was manually checked and where necessary corrected.

The corpus described in the previous paragraph contains 94,783 disyllabic word tokens, 11,482 of which have the RF tonal pattern.[9] These 11,482 tokens represent 707 orthographic word types. However, more than two-thirds of the tokens belong to one of five types: *ran2hou4* 'and then', *shi2hou4* 'during which', *bu2hui4* 'cannot', *hai2shi4* 'still', and *yi2yang4* 'likewise'. In order to avoid model predictions being heavily biased toward these five high-frequency words, we randomly sampled 300 tokens for each of these five types for inclusion in our data set.[10] We also excluded types with fewer than twenty occurrences from the data set, in order to avoid overfitting to low-frequency words. As a consequence, our initial data set comprised 4,516 tokens across fifty-three word types.

We extracted the sound files for our 4,516 tokens and measured their f0 values using the method described in Section 2.1 for the toy data set. For speakers self-identified as male, we set the pitch floor and ceiling to 50 Hz and 400 Hz, respectively. In order to make sure every token had sufficient data points for model fitting, we excluded extremely short tokens with fewer than six data points, which constituted about 5% of the data. Next, we removed tokens where an f0 extraction error was likely. Such errors usually result from pitch halving or doubling and lead to abrupt big changes in the recorded f0 values. We therefore first obtained, for each token, all of the f0 differences between consecutive measurements, and then calculated the standard deviation of these difference values. The standard deviation is large when f0 measurements are discontinuous and fluctuate abruptly. Tokens with standard deviations greater than the ninth decile of the distribution were considered to be outliers, hence likely to involve extraction errors, and were excluded from further analyses. Finally, two words were excluded because all of their tokens were contributed by a single speaker.

The final data set for the first analysis, reported in Section 2.6 below, consists of a total of 3,778 tokens representing fifty-one word types. Since these types do not include any heterographic homophones, there is a one-to-one correspondence between the orthographic labels of the tokens in our data and their canonical spoken forms. We therefore assume that tokens with the same label bear some similarity to one another in both form and meaning, that is, belong to the same word type. However, because these tokens were extracted from spontaneous speech, not every speaker produced every type. To ensure that no word was completely nested under speaker or vice versa, we checked that every type was produced by multiple speakers and that every speaker produced multiple types. In the data set of 3,778 tokens, the median number of speakers per word type is twenty, the mean is 24.45, and the range is five to fifty-two. The median number of word types per speaker is twenty-three, the mean is 22.67, and the range is twelve to thirty-five.

In the data set described above, the f0 values are positively skewed. We therefore log-transformed them to create a response variable with a distribution closer to Gaussian, a requirement for modeling with Gaussian GAMs.[11] Note that despite voiceless gaps in the data for certain tokens, the overall distribution

[9] These tokens also include tone sandhi words containing *yi1* and *bu4*, which have T2 realizations when followed by T4 syllables.

[10] The figure of 300 was an arbitrary choice at the upper end of the frequency range for the other words in the data set.

[11] For studies of auditory comprehension, several psychoacoustic scales are available, such as the Bark scale, that directly relate to human perception of pitch. However, as the interest of the present study is primarily in the pitch that speakers produced, we wanted to stay as close as possible to the physical signal and therefore simply log-transformed the Hertz values.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

of data points from all tokens combined is dense right across the [0,1] interval of normalized time. The model is therefore able to make accurate predictions for all tokens within this range, including those with voiceless segments, albeit with variations in confidence, as described for the toy data set in Section 2.1.

2.3. Predictors

The predictors used in our GAM models are described below. In addition to the core predictors related to our hypothesis, we also included, as far as possible, all of the variables that have previously been shown to influence tonal realization, as outlined in Section 1. These control predictors are grouped into three major categories: speaker-related, context-related, and segment-related.

2.3.1. Core predictors

- **Word type.** We coded each word token in our data set for its word type (**word**), using the orthographic representation of the token in the corpus as the identifier of its word type.[12]
- **Sense.** Unlike heterographic homophony, homographic homophony and polysemy are common in Mandarin disyllabic words. In lexicography, such diversity of meaning is usually addressed by attempting to enumerate the various possible senses of a given orthographic form. Similarly, in computational semantics, systems have been devised for disambiguating word senses from among a finite set of possibilities. The validity of this approach has been questioned, for example, by Kilgarriff (2007:29), who pointed out that there are 'no decisive ways of identifying where one sense of a word ends and the next begins'; polysemy is actually much more subtle and nuanced than a set of discrete possibilities would suggest. Nevertheless, sense annotations do capture, however crudely, some aspects of the variability in words' meanings. Furthermore, within the context of modeling pitch contours with GAMs, discrete senses are convenient because we can straightforwardly estimate specific pitch contours for each sense. We therefore coded every word token in the data set for **sense**, using a word sense disambiguation system (Hsieh & Tseng 2020) based on the Chinese WordNet (Huang et al. 2010). The possible values of this variable correspond to the senses identified by the disambiguation system. More than one sense was identified for thirty-five of the fifty-one words in the data set, with a total of 130 senses overall. All but two of the words had between one and five senses; of the two outliers, one had six senses and the other had nine. Note that, because the sense labels in our data are nested under the orthographic form and there are no synonyms, sense includes all of the information in word, plus additional information about the meaning of any given token.
- **Normalized time.** The points in time at which f0 measurements were taken were, for each token, transformed into a normalized time scale of 0–1 to produce the variable **time**.

2.3.2. Speaker-related controls

- **Gender.** Speakers self-identified as female usually have a higher pitch register and wider pitch range than speakers self-identified as male. Furthermore, with respect to tonal realizations in Taiwan Mandarin, a number of studies have documented detailed gender-dependent differences in various sociolinguistic domains (Fu 1999, Wu 2003, Huang 2008, Wu 2009). We therefore included **gender** as a simple control variable to account for intrinsic pitch height and range differences between speakers of different genders, as labeled in the corpus, and also allowed gender to interact with time, to accommodate possible gender-specific modulations of the pitch contour.

[12] Note that, although word frequency is an important predictor for response variables such as spoken word duration, it is not a factor that has been widely reported to co-determine the shape of pitch contours in Mandarin (but see Bi et al. 2015). We verified for our data that when word frequency is included in a model that also has access to word type, frequency is not significant, whereas word type is well supported. In this study, frequency of use is therefore not discussed any further. For further discussion of frequency and Mandarin tone for monosyllabic Mandarin words, see Jin et al. 2025.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

- **Speaker identity.** Speaker identity was included to account for any idiosyncratic tonal realizations specific to individual speakers. We included **speaker** not only as a main effect, but also in interaction with time, using by-speaker factor smooths.

2.3.3. Context-related controls

- **Duration.** The shapes of tonal contours are influenced by the time available to articulate them. Under time pressure, that is, when words are spoken quickly, speakers do not have enough time to realize the full tonal contours, which are therefore more likely to undergo reduction (Cheng & Xu 2015, Tang & Li 2020). We measured the duration of each token in seconds. As the distribution of these measurements was heavily skewed to the right, we log-transformed them before conducting GAM analyses. In our models, the variable **duration** is the log-transformed token duration.[13]
- **Adjacent tones.** When a tone is expected to start at a different pitch from where the previous one ends, for example, a falling tone followed by a high level tone, the degree of coarticulation, and hence deviation from the canonical tonal shapes, will be greater than when two tones are contiguous, for example, a high level tone followed by a falling tone (Shih 1988, Xu 1994). In addition, although the details differ across studies, tonal coarticulation is usually found to be bidirectional, that is, both anticipatory and preservatory (Shen 1990b, Xu 1997, Huang & Chiu 2023). For our analyses, we therefore coded the tonal category of each token's preceding and following syllables in the corpus. When a target token occurred utterance-initially or utterance-finally, the preceding or following tonal category was coded as 'null'. This gave us six possible tonal categories for both the preceding syllable and the following syllable: four lexical tones, one neutral tone, and 'null'. We therefore created the factor **adjacent_tone** with thirty-six levels to account for each possible combination of properties of the preceding and following syllables.
- **Utterance position.** The realization of tone in an utterance is also influenced by sentence intonation (Ho 1976, Tseng 1981, Shen 1989, 1990a). For example, statement intonation is often characterized by a downward trend, resulting in pitch declination (Shih 1997). Question intonation, by contrast, can potentially lead to a final rise, although this largely depends on the syntactic structure and/or emotive force of the question concerned (Lee 2005, Chuang et al. 2007). For the current study, we simply calculated the normalized position of each token in the relevant utterance. We defined an utterance as a sequence of words preceded and followed by a perceivable pause (regardless of duration), as indicated by the labels provided in the corpus. The variable **utterance_position** is the position at which a given token occurs in an utterance divided by the total number of words in that utterance. This predictor is therefore bounded between 0 and 1. For utterances with only one word, the utterance position was set to 1.[14]
- **Bigram probability.** Bigram probability is a measure of a word's contextual predictability based on its relative frequency of cooccurrence with the other words in its context; the higher the bigram probability, the more predictable a target word is in the given context. It has been found that a word's phonetic realizations are intimately related to its contextual predictability. In general, higher predictability is associated with shorter word duration and a greater degree of spectral reduction (Bell et al. 2003, Gahl et al. 2012). Specifically for tonal realizations in Mandarin, there is some evidence that these too are sensitive to contextual predictability; when a word is more contextually predictable or represents given information, its f0 excursion and range are found to be diminished (Hsieh 2013, Ouyang & Kaiser 2015). In the present study, following Gahl et al. 2012, we calculated the bigram probabilities of target

[13] In our data, token duration is strongly correlated with speech rate ($r = -0.57, p < 0.0001$). In other words, these two variables could not meaningfully be included as predictors in the same model (cf. Section 2.4). Since duration turned out to predict f0 better than speech rate did, we report models with duration.

[14] In many languages, f0 at the start of an utterance is related to the length of the utterance, with higher initial f0 for longer utterances. We therefore tried including utterance length in our models. Although overall model fit improved, the concurvity score of utterance length was as high as 0.84, suggesting that a large portion of the effect can be explained by other factors in the model (cf. Section 2.6). In particular, utterance length is negatively correlated with token duration ($r = -0.23$), so keeping both variables would have rendered their effects uninterpretable. We chose to remove utterance length and keep duration.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

tokens in two ways: **bigram_previous**, based on the preceding word, and **bigram_following**, based on the following word. These two variables are defined, respectively, as in 3 and 4, where $P(w_n|w_{n-1})$ is the probability of a word occurring given the previous word, $P(w_n|w_{n+1})$ is the probability of a word occurring given the following word, and Freq denotes word frequency in the corpus of Taiwan Mandarin.

(3) bigram_previous: $P(w_n|w_{n-1}) = \text{Freq}(w_{n-1}, w_n)/\text{Freq}(w_{n-1})$

(4) bigram_following: $P(w_n|w_{n+1}) = \text{Freq}(w_n, w_{n+1})/\text{Freq}(w_{n+1})$

2.3.4. Segment-related controls

- **Vowel height.** It has long been recognized that different vowels have different intrinsic pitch, a finding established for a great number of different languages, including Mandarin (Ho 1976, Ladd & Silverman 1984, Shi & Zhang 1987, Whalen & Levitt 1995). Specifically, high vowels tend to have higher f0 values than low vowels. For our disyllabic words, we coded the vowel heights of the vowels of the first and second syllables as two separate predictors, **vowel1** and **vowel2**, respectively. For monophthongs such as /i/ and /a/, we distinguished between three vowel heights: 'high', 'mid', and 'low'. For diphthongs such as /aɪ/ and /eɪ/, which are characterized by within-vowel changes in height, we added two additional levels: 'low-high' and 'mid-high'. This means that there are theoretically twenty-five possible combinations of vowel1 and vowel2. Our data set included twenty of these.
- **Onset.** The effect of onset consonant on f0 has been studied in considerable detail in Mandarin. Ho (1976), for example, found that after voiced consonants, pitch tends to start lower than after voiceless consonants. For stops, aspiration also results in lower initial f0, although the magnitude of this effect appears to be tone-dependent (Xu & Xu 2003). Following Howie (1974), we distinguished onset types according to manner of articulation, voicing, and aspiration. For each of the two syllables in our target words, we distinguished between 'aspirated-affricate', 'aspirated-stop', 'unaspirated-affricate', 'unaspirated-stop', 'voiceless-fricative', and 'voiced'. Syllables that do not have onsets and that instead start with vowels or glides were coded as 'null'. Our data set contained thirty different combinations of the onset type of the first syllable (**onset1**) and that of the second syllable (**onset2**).
- **Rhyme structure.** Although effects appear to be unstable and are not always reliably observed, some studies have reported variation in f0 for different Mandarin syllable types (Howie 1974, Xu 1998, Fon & Hsu 2007). In our models, we therefore included a control variable for syllable structure. Given the strict phonotactic constraints governing the syllables of Mandarin, a syllable can maximally be composed of an onset consonant, a prenuclear glide, a nucleus vowel, and finally a coda consonant (Duanmu 2007). In some theoretical descriptions, a coda consonant must be a nasal; in other descriptions, it can be either a nasal or a postnuclear glide, as in the case of /aɪ/, for example. In this study, we coded the latter cases as diphthongs in the vowel height predictors, vowel1 and vowel2, and coded for a coda consonant only when the syllable included a final nasal. For each of the two syllables in our target words, we therefore coded the structure of the rhyme as 'V', 'GV', 'VN', or 'GVN'. This coding specifies, for a given syllable, whether there is a prenuclear glide, as well as whether there is a final nasal. Applied to the two syllables of our target words separately (**syllable1** and **syllable2**), we obtained fourteen attested combinations of rhyme structures.

2.4. Modeling strategy

Because pitch typically changes continuously and gradually across the time course of an utterance, it is inevitable that our response variable of f0 measurements is characterized by significant autocorrelation. That is, the f0 at time t is correlated with and can thus to some extent be predicted from the f0 at $t - 1$. Autocorrelation is particularly problematic for regression modeling because the residuals of an autocorrelated response variable are also unavoidably autocorrelated. This means that a central assumption of regression modeling is violated, namely, that the residuals should be independent of one another.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

In GAMs, the issue of autocorrelation can be addressed by incorporating a first-order autoregression model for the errors, denoted as AR(1). An AR(1) model is a linear model that predicts a given value of a time series from the immediately prior value; including an AR(1) process in a GAM enables the model to accommodate structure in the residuals by positing a linear relationship between a given residual and its preceding residual. Specifically, the residual at time t is modeled as the sum of a proportion ρ of the residual at time $t - 1$ plus Gaussian noise. In all of the models presented below, the AR(1) process is included. In order to determine the appropriate value of ρ, we first fitted a model without AR(1), then calculated the autocorrelation of the residuals in this model at lag 1, that is, comparing the residual at each time point with the immediately preceding residual. Finally, we fitted the model again, now with AR(1), for which we set ρ to the lag 1 autocorrelation.[15]

A standard method for evaluating the effects of different predictors in regression modeling is to compare nested models in which individual variables are progressively added or removed. However, this approach assumes that the predictors are relatively independent of one another, so that the contribution of each predictor to overall variance can be identified. If highly correlated predictors are entered into a single model, it becomes impossible to separate the effects of one from another and hence to assess their relative contributions. It is self-evident that the segmental makeup of a word cannot be independent of that word and that, in our models, the core predictor word is almost completely identified by the segment-related controls and vice versa. Thus, in a model that included both word and the controls, it would be impossible to evaluate the effects of the different predictors. The same applies to word and sense, since the latter is completely nested under the former. Consequently, the standard approach to model comparison was not available to us as a means of testing our predictions, since we were specifically aiming to tease apart the effects of different variables.

Instead of comparing nested models that would have been linguistically uninterpretable, we adopted the following two-fold modeling strategy. First, in order to evaluate whether a given predictor was relevant for understanding f0 contours, we made use of Akaike's information criterion (AIC; Akaike 1998). AIC estimates the relative quality of alternative models for a given set of data, by balancing goodness of fit against the number of parameters for each model, penalizing excessively complex models that might overfit the data. The method can be used for nonnested models and is therefore well suited to our analysis. The best model is the one with the lowest AIC value.[16] Second, we investigated the adequacy of our predictors by means of cross-validation. That is, we held out a small portion of our data set as testing data, fitted our models to the remaining data, and then assessed model accuracy on the testing data. We repeated this process for 100 different data splits selected by stratified random sampling such that every word type was represented in both the training and test data, in similar proportions. In other words, the test sets consisted entirely of novel tokens but no novel types, simulating the situation for human language use with previous experience. Using cross-validation, we could directly assess the precision with which our models predicted novel, previously unseen, data and establish whether inclusion of a predictor improved prediction accuracy.[17]

We used this modeling strategy to explore the first two predictions presented in Section 1, repeated here for convenience:

(i) Word type will be a stronger predictor of tonal realization than all of the previously established word-form-related predictors combined.

[15] We note that the current implementation of AR(1) in the mgcv package is suboptimal, as it allows for only one ρ value per model, effectively treating the entire data set as one big time series. However, the degree of autocorrelation actually varies considerably across the tokens in our data set. Applying a single ρ correction therefore undercorrects for some tokens but overcorrects for others, and may in part give rise to residuals that do not approximate a normal distribution and cannot be corrected to a normal distribution. We discuss this further in the supplementary material, available at https://osf.io/nwv74/.

[16] For model comparison using AIC, the evidence ratio is calculated as $e^{\frac{\Delta AIC}{2}}$. A decrease of ten AIC units, for example, means that the model with the smaller AIC is 148.4 times more likely than the alternative model to generate the observed data.

[17] The cross-validation approach we used is sometimes called repeated training/test split or Monte Carlo cross-validation. It is one of the most common cross-validation methods in machine learning (Zhang 1993, Kuhn & Johnson 2013).

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

(ii) Information about a word's meaning in context will improve prediction of its tonal realization, compared with prediction based on word type alone.

Prediction (i) is based on the hypothesis that the unique pitch contour of each spoken Mandarin word token is determined in part by the meaning of that token. If this is correct, then variations in tonal realization arise not only from the mechanical constraints on articulation previously described in the literature, but also as a result of form-meaning connections in the lexicon. We therefore investigated whether using word as a predictor would lead to a more precise model of tonal realization, as compared to a model using the set of segment-related predictors (e.g. vowel height) that have previously been identified as relevant. Since word incorporates not only the segmental form of a word type but also the associated semantics, our expectation was that it would be a superior predictor compared to all of the previously identified segment-related variables considered jointly, when we controlled for contextual effects such as token duration and adjacent tones.

As a first step, we fitted a baseline GAM to logarithmically transformed f0, including all of the speaker-related and context-related control variables described in Section 2.3, in interaction with time. We then fitted six further models, each with one segment-related control variable added to the baseline, allowing us to investigate the articulatory effects described in the literature. To compare these effects jointly against the effect of word, we fitted two additional models: one with all six segment-related control variables included in addition to the baseline, and the other with only word as an additional predictor. Both word and the segment-related controls (vowel1, vowel2, onset1, onset2, syllable1, syllable2) were modeled as factor smooths in interaction with time; in other words, all of these variables have both random intercepts and random wiggly effects on the shapes of the predicted contours.

Complex models, such as our GAMs, tend to raise a concern that they may be overfitted, that is, fit the training data so closely that the model fails to generalize to new data. Since prediction (i) concerns the effect of word type, we took particular care to check that any effect of word was not attributable to overfitting. The first step in this process was the cross-validation procedure described above, which evaluated the model's generalizability; we then backed this up using a slightly different set of data. In addition, we conducted two further analyses designed to address the two main potential sources of overfitting in our model, namely the complexity of the model overall and the smoothing parameter of the factor smooth for word. These extra measures are described in the following paragraphs.

First, we repeated the GAM analysis using a slightly different data set. Recall from Section 2.2 that we had randomly selected 300 tokens for the most frequent word types in our data to avoid the analysis being biased toward these types. The four most frequent types each had enough tokens in the corpus for us to be able to completely replace their tokens in the data set; these were *ran2hou4* 'and then', *shi2hou4* 'during which', *bu2hui4* 'cannot', and *hai2shi4* 'still'. We therefore created a new data set consisting of different tokens of each of these four types (300, 300, 217, and 219 new tokens, respectively), along with the original tokens of all of the other types. We then conducted the same GAM analysis, fitting a model with all of the speaker-related and context-related control variables as well as the factor smooth for word, in interaction with time. If the original analysis resulted from overfitting, using different tokens would change the results.

Next, using the original data set, we investigated the relationship between model complexity and the effect of word. To do this, we ran two simplified models, each with most baseline predictors removed. The first model included only one speaker-related control, gender, and one context-related control, duration, alongside the core predictors word and time. The second model added speaker as an additional predictor. If the effect of word in these simplified models was found to closely resemble its effect in the full model, this would suggest that its contribution to explaining variation in f0 is not dependent on interactions with other variables. Such consistency across model specifications provides strong evidence of a predictor's robustness and reliability as a main effect.

Finally, we assessed the quality of the smoothing parameter of the word factor smooth term in our model. When a GAM is fitted, the smooth for any given predictor is assigned a smoothing parameter that controls the degree of wiggliness of the predicted curve. In selecting parameters, the GAM seeks a balance

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

between staying faithful to the data by keeping residuals as small as possible and keeping the model as simple as possible by penalizing for wiggliness. If the parameter is set too high, wiggliness is strongly penalized, and the resulting curve might be so smooth that the model misses important patterns; for the highest values, the model simply predicts a straight line. By contrast, if the parameter is set too low, penalization of wiggliness is very mild, and the model runs the risk of overfitting, that is, staying so faithful to the data that the predicted curve is excessively wiggly and captures noise instead of the true underlying trend.

The mgcv package (Wood 2017), employed in our modeling, uses a mathematically robust method to estimate smoothing parameters by evaluating a range of candidate values and selecting the value that, for the data set in question, achieves the best balance between capturing significant nonlinear patterns and avoiding overfitting to small fluctuations in the data. Although this empirical Bayes method is mathematically well justified, the reader may wonder how robust the method is for complex models with many complex smooth terms, such as those used in the present study. Since it is possible to manually select a smoothing parameter, we therefore generated a series of GAM models in which we systematically varied the smoothing parameter for the word factor smooth to identify the optimal value, that is, the value yielding the lowest prediction error on held-out data. We then compared this value with the automatically selected smoothing parameter for the word factor smooth in our final model. A parameter value close to the manually identified optimum would constitute evidence that the model was not overfitted.

If prediction (i) is correct, and we find an effect of word type that cannot be attributed to overfitting, this will provide some support for the hypothesis that meaning contributes to the realization of tone. However, it will not enable us to draw any firm conclusion. This is because word type subsumes information about form as well as meaning. Even if word performs better than the combination of all previously identified segmental predictors, we cannot be certain that this difference is due to meaning; it is possible that some additional aspect of word form influences tonal realization but has not yet been identified as a predictor. However, at the word level, it is impossible to tease these effects apart. Since all of our word types have the same RF tonal pattern, if we were to enter every detail of a token's segmental makeup into the model, we would effectively be specifying its word type. Prediction (ii) addresses this issue by separating variation in meaning from variation in form.

Prediction (ii) is based on the hypothesis that variation in tonal realization is partly determined by token-level variation in meaning, that is, variation in meaning within word type. To address this prediction, we compared the model with a factor smooth for word as the sole predictor added to the baseline against a model in which a factor smooth for sense was the sole predictor added to the baseline, both in interaction with time. Recall that, because the sense labels in our data are nested under the orthographic form, sense includes all of the information in word, plus additional information about the meaning of any given token. If semantics is indeed at issue, replacing word by sense should therefore improve model fit even further.

In the data set for the word-type analysis, the frequency distribution of sense is skewed toward the right, with about half of the senses having no more than thirteen tokens. To make sure that all senses included in the models had sufficient tokens for statistical evaluation, we therefore used only a subset of the data for the models used to investigate the effect of sense. Since the median number of tokens per sense in the data set was 13.5, we excluded senses with fewer than fourteen tokens. This left us with a data set of 3,458 tokens representing sixty-five senses across a smaller set of forty-eight word types. We used this smaller data set for models evaluating sense as a predictor.[18] The statistical analysis proceeded in the same way as described above for word type, except that we fitted an additional model with a factor smooth for sense as the only predictor in addition to the baseline.

[18] For completeness, we note that thirty-five words in the smaller data set have only one sense. Of these thirty-five words, six have only one sense listed in the Chinese WordNet, and seven are not included in the vocabulary of the Chinese WordNet. For the rest, either the tagger identified only one sense, or only one sense of the word had more than thirteen tokens in our data set.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

2.5. The baseline GAM

We fitted the baseline GAM to log-transformed f0 (pitch), using the following model specification.[19]

(5) $\text{pitch} \sim \text{gender} + s(\text{time}, \text{by} = \text{gender}) +$
$s(\text{time}, \text{speaker}, \text{bs} = \text{'fs'}, \text{m} = 1) +$
$s(\text{duration}) + ti(\text{time}, \text{duration}) +$
$s(\text{utterance_position}) + ti(\text{time}, \text{utterance_position}) +$
$s(\text{bigram_previous}) + ti(\text{time}, \text{bigram_previous}) +$
$s(\text{bigram_following}) + ti(\text{time}, \text{bigram_following}) +$
$s(\text{time}, \text{adjacent_tone}, \text{bs} = \text{'fs'}, \text{m} = 1)$

The first line of this model specifies a main effect for gender, in order to account for male voices being lower on average than female voices. We used treatment (i.e. dummy) coding, with female as the reference level. In addition, the first line includes a separate smooth for each gender. The upper left-hand panel of Figure 3 plots the predicted contours for speakers self-identified as female and those self-identified as male. Similar to the pattern observed in read speech (cf. Figure 1), the realization of the RF tonal pattern in spontaneous speech is characterized by a shallow fall, followed by a long rise, and finally a much larger fall. Speakers self-identified as male show reduced pitch excursion compared to those self-identified as female, presumably due to male voices having a more compressed pitch range. The second line of the model specifies by-speaker nonlinear random effects, using factor smooths. These factor smooths specify, for each speaker, the specific way in which that particular speaker modulates the general f0 contour associated with their gender.[20]

The next four lines in the model formula deal with the four numerical context-related controls, namely duration, utterance_position, bigram_previous, and bigram_following. For each of these variables, the model includes a main effect smooth in combination with a tensor product smooth for the interaction of the given variable with time (using an interaction-specific tensor product smooth specified with 'ti'). The upper mid panel of Figure 3 plots the modulating effect of token duration on the base contour of speakers self-identified as female. Shorter duration, represented by darker lines, reduces the amplitude of the wave. The effect of position in the utterance is depicted in the upper right-hand panel of Figure 3. The tonal shape is clearly most different when the word occurs toward the end of an utterance, in which case we observe an earlier peak. This might be due to the fact that we coded words in singleton utterances as occurring at the end of the utterance. The left and mid panels in the lower row of Figure 3 present the effects of the bigram probabilities given the preceding and following word, respectively; higher bigram probabilities are represented by lighter shades of gray. When bigram_previous is high, meaning that the word is more expected given the preceding word, the f0 excursion is reduced. This effect parallels the finding of Hsieh (2013) that f0 excursion in Taiwan Mandarin is diminished in conditions of high semantic predictability. The effect of bigram_following in our model is much smaller than the other contextual effects but appears to go in the opposite direction from bigram_previous, with higher values associated with sharper peaks in f0.

The final line of the model specifies factor smooths for adjacent_tone, requesting a separate smooth for each of its thirty-six levels. The effect of adjacent_tone is presented in the lower right-hand panel of Figure 3 for those tokens that have T1 as following tone. Unsurprisingly, the four predicted contours end

[19] All model formulae and summaries are also provided in the supplementary material: https://osf.io/nwv74/.

[20] The syntax 'bs = "fs"' on the second line of the baseline model specification in 5 has a similar effect to the 'by' argument on the first line; both terms request a smooth for each level of a single factor variable (in this case, speaker and gender, respectively). The 'by' argument is generally preferred when the factor has few levels and the levels are of interest, as in the case of gender. For computational efficiency, the 'fs' argument is preferred when there are many levels and these levels are of less direct interest, as in the case of speaker. When specifying a by-smooth, a separate term requesting a main effect for the intercept needs to be specified. In contrast, a factor smooth incorporates adjustments to the intercept, thus effectively calibrating the individual smooths for their relative position with respect to the general intercept. See Baayen et al. 2022 for detailed discussion.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

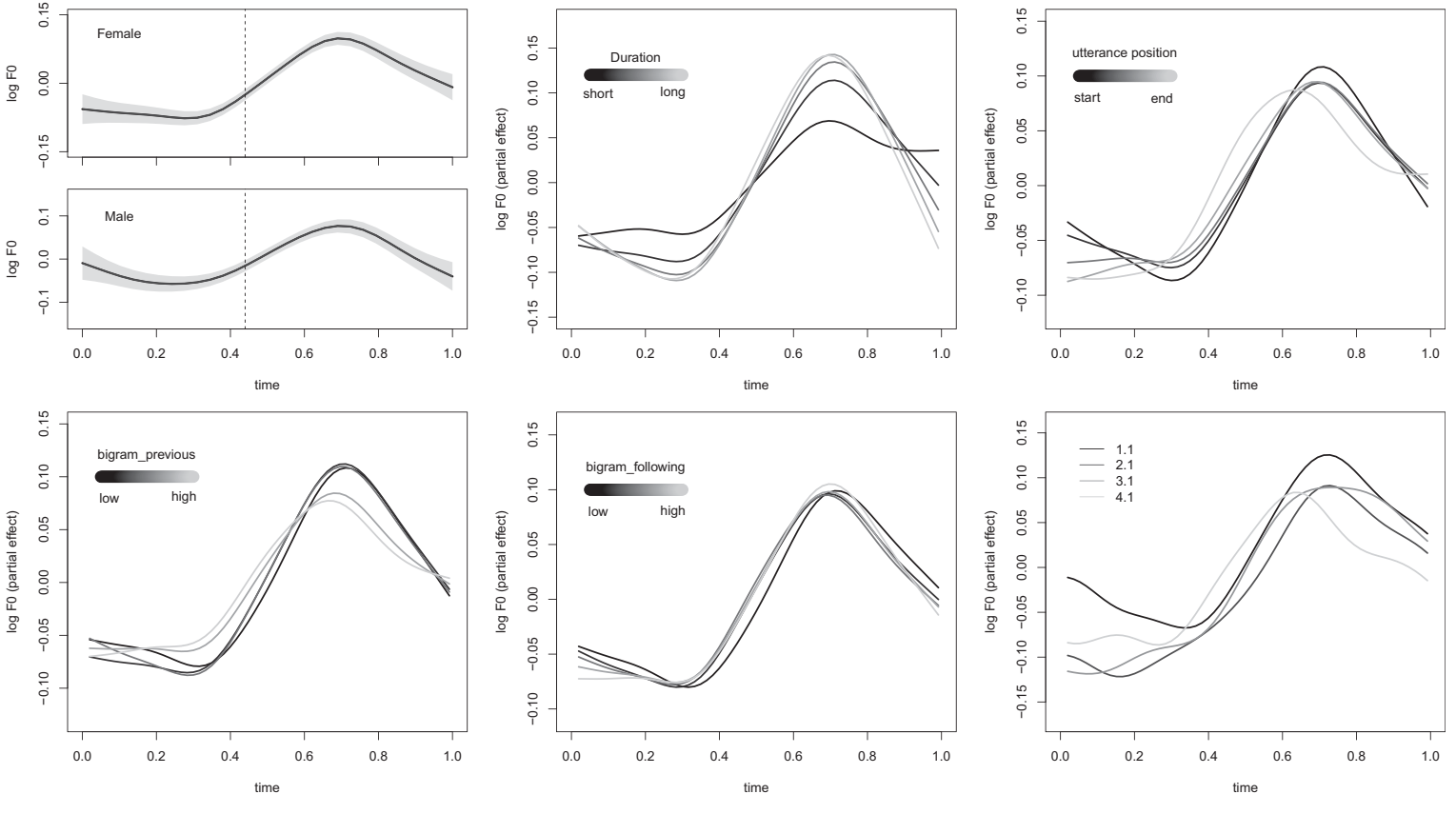


Figure 3. *Partial effects in the baseline GAM. The upper left-hand panel shows the predicted base contours for speakers self-identified as female and speakers self-identified as male. The next four panels show, for female speakers, how the base contour is modulated by duration, utterance position, previous bigram probability, and following bigram probability, respectively. The final panel presents, again for female speakers, the effect of tonal coarticulation with the tone of the preceding word, when the following word has a high-level tone.*

similarly. However, the initial part of the contour diverges considerably, depending on the preceding tone. As expected, tonal context has a very large effect on the shape of the f0 contour.

In what follows, we take this model as our baseline, with all contextual covariates controlled for, and compare the effects of predictors representing individual aspects of word form with the effect of word type as a whole.

2.6. Results and discussion: word type

2.6.1. Evaluation of predictors

The left-hand panel of Figure 4 presents the relative improvement in model fit, as compared to the baseline model, for the six models with an additional factor smooth for a single segment-related control, the model with factor smooths for all segment-related controls (henceforth the omnibus-segment model), and the model with only a factor smooth for word as additional predictor. Improvement in model fit is gauged by the magnitude of any decrease in AIC. As can be seen, each individual segment-related control improves model fit, a result that dovetails well with the previous studies summarized in Section 1.

In the omnibus-segment model, the inclusion of all six segment-related controls jointly provides a substantially better fit than any single predictor and leads to a fall in AIC of 4,938 units compared to the baseline model. However, this is a very poor model from a statistical point of view because its key predictors are correlated with one another. In a GAM, the **concurvity** score of a predictor is a number bounded between 0 and 1 that measures the degree to which the effect of a given independent variable can be predicted by one or more of the other independent variables in the model.[21] If a predictor's concurvity is low, this predictor has its own explanatory value; however, if the concurvity is high, the predictor is strongly confounded with other predictors. The light gray dots in the right-hand panel of Figure 4 represent the concurvity scores of the segment-related controls in the omnibus-segment model. It can be

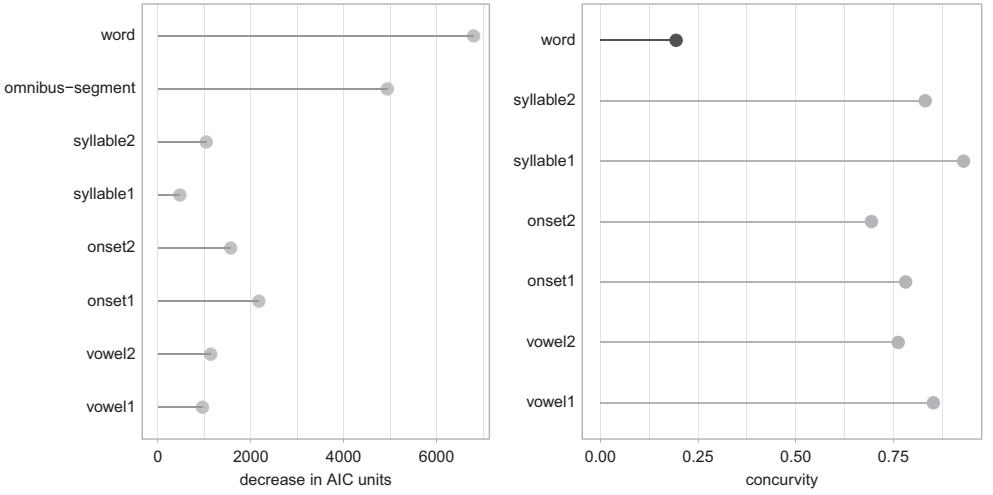


Figure 4. *The left-hand panel shows model fit improvement gauged by decrease in AIC units when a predictor (or set of predictors) is added to the baseline model for the word-type analysis. The right-hand panel shows the concurvity score of individual predictors in two models using the full data set of 3,778 tokens: the omnibus-segment model (light gray), with factor smooths for all segment-related control variables added to the baseline, and the word model (dark gray), with only a factor smooth for word added to the baseline.*

[21] Concurvity is the nonlinear equivalent of collinearity.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

seen that the concurvity scores of all predictors are high, indicating that the effects of the segment-related controls are confounded with one another, rendering interpretation of the individual effects difficult, if not impossible.[22] Despite the fact that the omnibus-segment model is linguistically uninterpretable, we have presented it here purely for comparison with the model with only word as additional predictor.

The concurvity score of the predictor word in the word-type model is represented by the dark gray dot in the right-hand panel of Figure 4. It can be seen that word has low concurvity with the other predictors in this model (i.e. the baseline controls), so that the interpretation of the effect of individual word types on the f0 contours is straightforward. Furthermore, adding only word to the baseline model results in an even better fit than the omnibus-segment model, with a fall of 6,795 AIC units compared to the baseline. Although the segment-related controls address all of the word-internal properties previously found to influence tonal realization, the contribution of just word by itself to the model fit is much stronger. The difference of 1,857 AIC units between the omnibus-segment model and the word model means that the probability of the word model giving a better fit approaches infinity. It is clear that the association between word type and tonal realization cannot be reduced to the segment-related constraints on articulation previously described in the literature. The actual pitch contour is richer than what can be predicted from all phonetic features that have been found to be relevant. Note, however, that we do not claim that the segmental predictors are irrelevant, as our analyses clearly demonstrate that they all improve on the baseline model. We also do not claim that once word is included, the segmental predictors are irrelevant. What we do claim is that word by itself outperforms all segmental predictors jointly, and hence has added predictive value over and above the segmental predictors.

The predicted pitch contours of a sample of fifteen word types are presented in Figure 5. To better visualize how the word-specific tonal modulations differ from one another, the partial effect predicted for each word has been added to the general contour for speakers self-identified as female (cf. Figure 3). In general, the fall-rise-fall pattern can be observed for all of these words, but the details of tonal excursions differ significantly from word to word. For example, while the initial falling part is very prominent for words like (c) *jue2ding4* 'decision' and (f) *quan2bu4* 'all', it is rather muted for (m) *yi2ban4* 'half' and (d) *nian2ji4* 'age'. In addition, in terms of the degree of undulation, some words have more reduced tonal range, such as (a) *bu2shi4* 'not' as compared to (j) *wen2hua4* 'culture', which has an extensive f0 excursion.

A closer inspection of Figure 5 reveals that, as expected, some of the word-specific contours appear to be consistent with the words' canonical segmental properties. For example, the initial fall appears to be more salient when the onset of the first syllable is a voiceless sibilant, as in (l) *xi2guan4* 'habit', or an affricate, as in (o) *za2zhi4* 'magazine', with the onsets /ɕ/ and /ts/, respectively. This pattern might be partly associated with the tendency for voiceless onsets to be followed by higher initial f0 in the following vowel, as compared to voiced onsets (Ho 1976). However, it should be kept in mind that the model is imputing f0 values for these voiceless onsets, where no periodic wave form is actually produced. In Figure 5, the plotted 95% confidence intervals for the early timesteps in these words are wide and can be seen to partially straddle the horizontal axis. Since the horizontal axis represents no effect, there is no good evidence for modulation of the general f0 contour early on in these words. Another segmental property that is to some extent visible in Figure 5 is the length of the second syllable, relative to the length of the first syllable. If the second syllable is relatively short as compared to the first syllable, for instance, (f) *quan2bu4* 'all' and (g) *rong2yi4* 'easy', the final fall tends to be attenuated, as expected given that relatively less time is available to physically implement a large fall in pitch. Nevertheless, the superior performance of word over the segment-related controls in our models suggests that such articulatory effects are not the only elements at play in determining tonal realization.

2.6.2. Cross-validation

Recall from Section 2.4 that the goal of the cross-validation analysis is to investigate whether the fitted model can make precise predictions for held-out data, that is, tokens that were withheld from the model

[22] As explained in Section 2.4, this high level of concurvity, which would be further increased by the addition of word, is the reason why we cannot create a meaningful model that includes both word and the segment-related controls.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

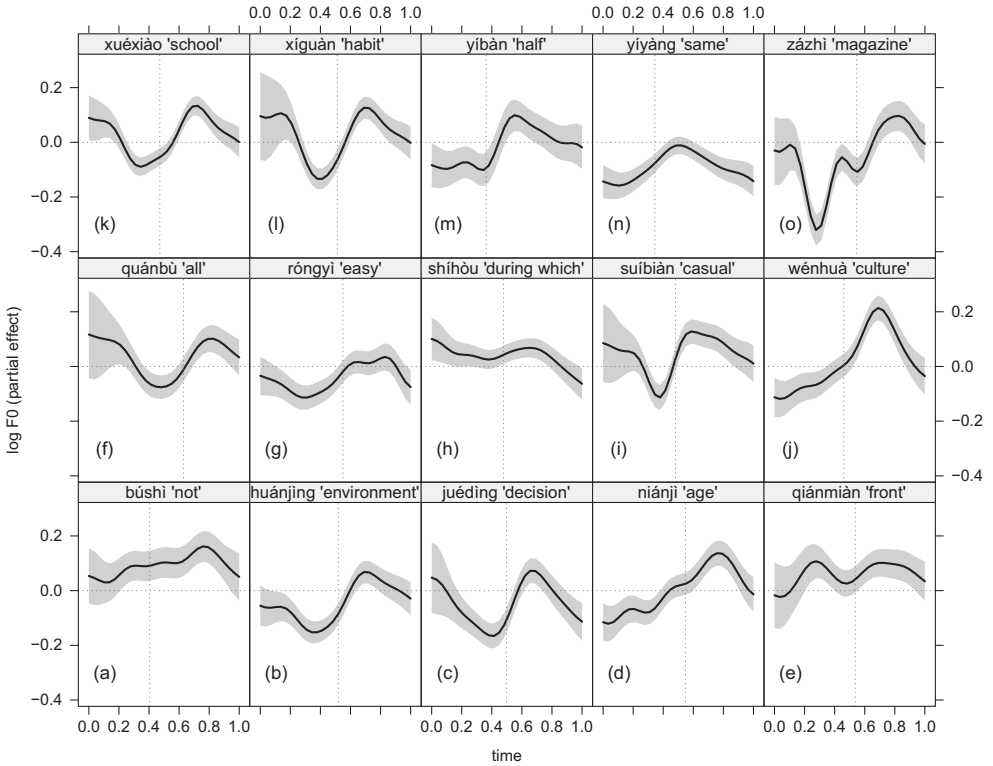


Figure 5. *Examples of the pitch contours predicted by the general smooth for time for female speakers, combined with the partial effects of the factor smooth for word. These partial effects do not include the general intercept or the differences in pitch between female and male speakers. As they represent the pure effect of word on the pitch contour, irrespective of other predictors, the curves are centered around the* y*-axis (indicated by a horizontal dotted line). The vertical dotted lines in the panels indicate the average (word-specific) syllable boundary.*

during model fitting (training). If our hypothesis is correct, a GAM that has access to word should provide superior prediction accuracy on held-out data compared to a GAM that has access only to the segment-related controls. We therefore evaluated prediction accuracy under cross-validation. We held out 10% of the current data as test data and used the remaining 90% as training data. Every word type was represented in both the training data and the test data, with approximately the same distribution in each set. We used stratified random sampling to produce 100 different splits with these properties.

We fitted ten models to the training data. In addition to the baseline model and the eight models assessed in Figure 4 (left-hand panel), we added one more model that was given data in which the values of word were randomly permuted. That is, tokens of a given word were now assigned different random word labels. In what follows, we refer to this model as the random-word model. If the effect of word is genuine, then random permutation of the word labels should substantially reduce prediction accuracy. To quantify model accuracy, we obtained the models' predictions for the f0 contours of the held-out test data and calculated the sum of squared errors (SSE) as a measure of prediction accuracy.[23] The SSE for the held-out data of a given model should be smaller than the SSE of the baseline model if the addition of one or more

[23] The sum of squared errors (SSE) is the sum of the squared difference between the observed and predicted values. A smaller SSE indicates more precise model predictions.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

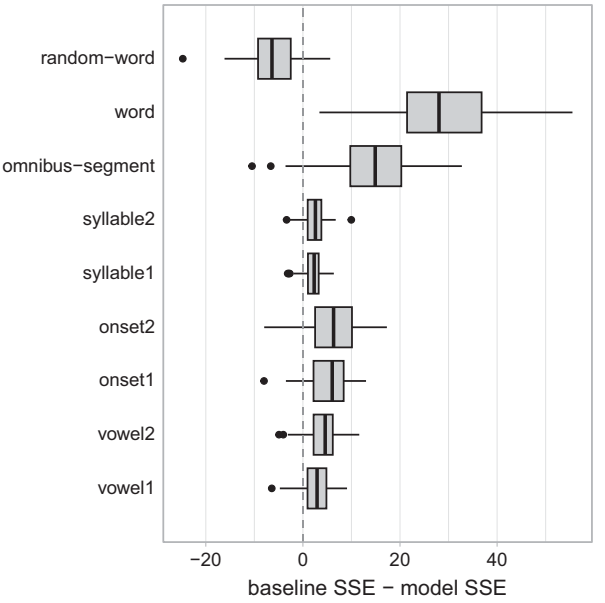


Figure 6. *Model accuracy under 100 runs of cross-validation for the word-type analysis. The boxplots represent the distributions of reduction in SSE.*

predictors indeed improves that model's prediction accuracy. We ran all ten models with all 100 random splits to cross-validate our results for the 100 held-out data sets.

Figure 6 presents boxplots of the SSE difference between the baseline model and each of the nine models of interest. Positive values indicate that the model in question offers more precise predictions than the baseline, as its SSE is smaller than the baseline's. All of the individual segment-related controls increase prediction accuracy over the baseline to some extent, albeit to varying degrees. However, the omnibus-segment model and the word model produce substantially greater increases in prediction accuracy, with the latter reducing the SSE to a larger extent than the former, replicating the model fit results. Moreover, when word labels are randomized, model accuracy plummets: the SSE of the random-word model is greater than that of the baseline model.

2.6.3. Model robustness

Figure 7 shows the pitch contours predicted by the general smooth for time for speakers self-identified as female, combined with the partial effects of the factor smooth for word, in models using two slightly different data sets, as described in Section 2.4. The upper panels show results for the four highest-frequency words, which are represented by completely different tokens in each data set. The lower panels show results for four randomly selected lower-frequency words, which have the same tokens in each data set. As can be seen, the two models predict very similar contours. The similarity is greatest for the lower-frequency words, as expected given that these are represented by exactly the same tokens in each case; but the similarity is also evident for the high-frequency words, where there is no overlap in the data. These results indicate that the word-specific pitch modulations observed with the original data set are not due to overfitting, which would prevent the model from generalizing to new data. On the contrary, very similar contours are observed across different samples of tokens for the high-frequency words, with almost no knock-on effect on the predicted contours for the low-frequency words.

Figure 8 shows the partial effect of the factor smooth for word, for the same eight words as shown in Figure 7, in three models of varying complexity. Model A includes only gender, duration, word, and time as predictors; model B includes the same predictors as model A, with the addition of speaker; model C is the full word model with all speaker-related and context-related controls, as well as word and time. It can

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

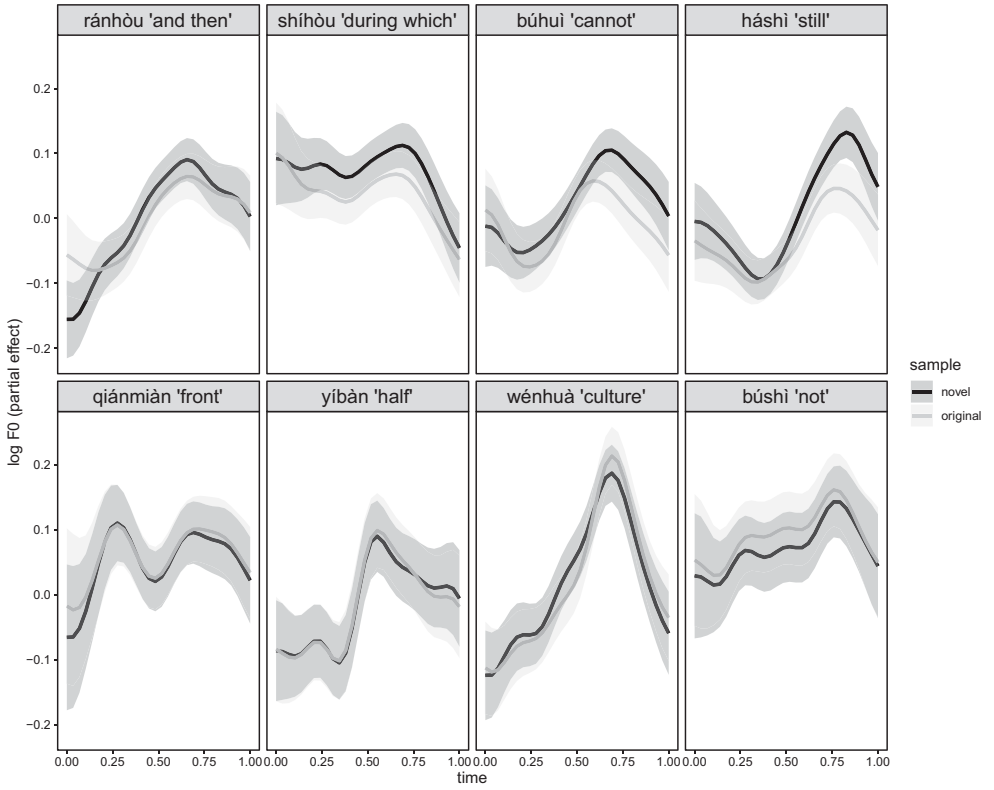


Figure 7. *Examples of the pitch contours predicted by the general smooth for time for female speakers, combined with the partial effects of the factor smooth for word. Predictions obtained with the novel and original data sets are indicated by dark and light gray, respectively. The upper panels present words that have different samples of tokens in the two data sets, whereas the lower panels present a random selection of four words of which the same tokens were used in the two analyses.*

be seen that the shapes of the curves are extremely similar in all three models. In other words, the shapes of the word-specific partial effects in the full model are not artifacts of the other predictors in the model; the word effect is there, irrespective of model complexity, and survives inclusion of all other predictors that are known to affect tonal realization.

Figure 9 shows the plots used to assess the quality of the smoothing parameter selected by the GAM for the word factor smooth. To produce these graphs, we created seven versions of our final word model, in which the smoothing parameter for the word factor smooth was manually set to its value selected by the GAM, multiplied by 0.001, 0.01, 0.1, 1, 10, 100, and 1000, respectively. We then split the data into 90% training data and 10% test data, and fitted all seven models to the training set, evaluating their performance on the test set. The process was repeated for thirty different training/testing splits. The left-hand panel of Figure 9 shows the results for the training data, averaged across the thirty runs, and the right-hand panel shows the equivalent results for the testing data. The horizontal axes represent values of the smoothing parameter for the word factor smooth; the vertical axes represent model performance, measured as the mean squared error (MSE) of predicted f0 relative to actual f0. For both training and testing, we find that the MSE changes across parameter settings in a U-shaped manner. At the lowest settings, the testing error is far greater than the training error, indicative of overfitting; the models with the smallest smoothing parameters do not generalize well and produce imprecise predictions for the test data. Conversely, at the highest parameter settings, the errors for both training and testing rise sharply, indicative of underfitting; these models are not sufficiently faithful to the training data to be able to capture nonlinear patterns. In this

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

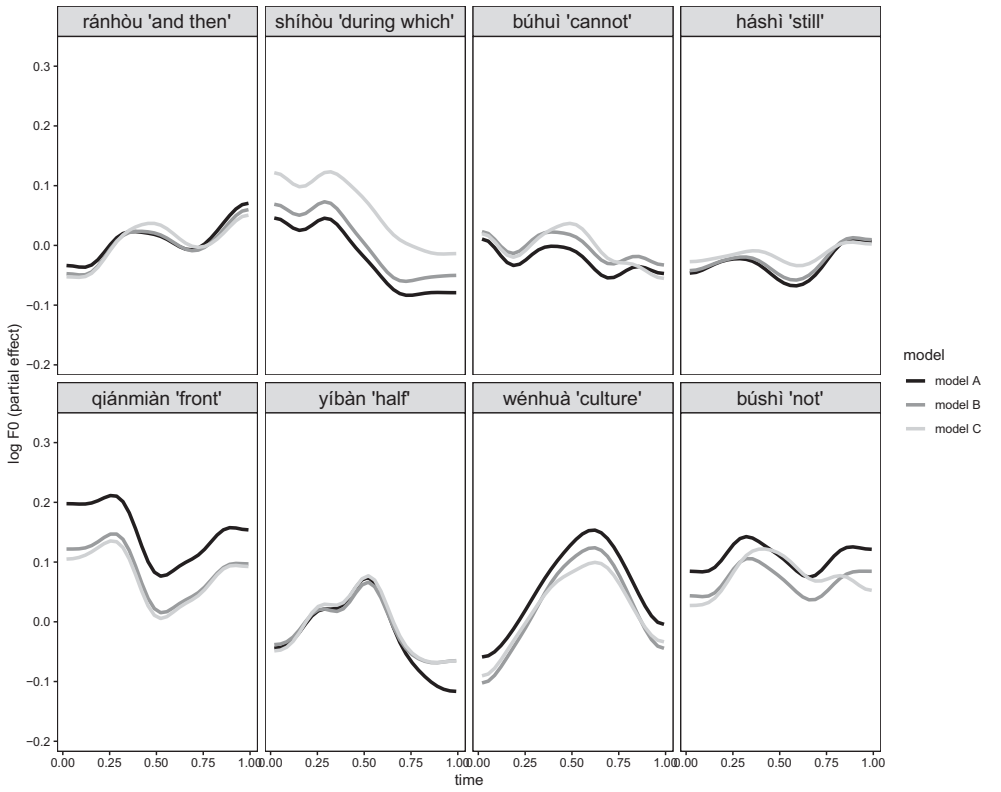


Figure 8. *The partial effect of the word factor smooth predicted by the three models for a selection of eight words.*

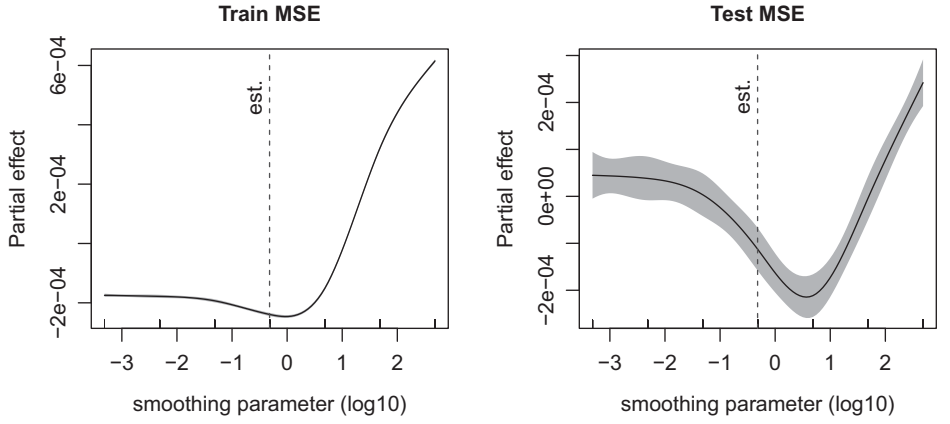


Figure 9. *The effect of smoothing parameters on the mean squared error (MSE) for training (left) and test (right) data. The dashed lines indicate the estimated smoothing parameter by GAM in the full model. For both curves, a 95% confidence interval is indicated, which for the training data is so narrow that it is hardly visible.*

analysis, the optimal value for the smoothing parameter is defined as the value that minimizes prediction error on the testing data, that is, the value corresponding to the lowest point on the testing error curve. For comparison, the smoothing parameter selected by the GAM is represented by the vertical dashed line on

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

each graph. The fact that this line falls only slightly to the left of the turning point in the testing curve shows that the model selects a parameter that appropriately balances faithfulness to the training data with generalizability. Nevertheless, as an additional check, we refitted the model with the smoothing parameter for the word factor smooth manually set to the 'optimal' value (0.56): the partial effect of word remained highly significant ($p << 0.0001$).

There are at least four possible reasons why the GAM selects a smoothing parameter for the word factor smooth that is slightly lower than the optimal value identified by our manual analysis. First, the use of a factor smooth assumes that all word-specific smooths are governed by the same smoothing parameter, which may be an oversimplification. Second, the GAM model has to estimate not one smoothing parameter, but many smoothing parameters. We have considered prediction accuracy for word, but not for speaker nor for any of the context-related controls. The requirement to find an overall optimum for multiple simultaneous constraints might mean that individual parameters deviate slightly from their optimal values considered in isolation. Third, as described in fn. 15, the implementation of AR(1) in the mgcv package is suboptimal for our data, as it assumes a constant level of autocorrelation, whereas the degree of autocorrelation actually varies considerably across the tokens in our data set. Finally, the residuals of the model depart markedly from a normal distribution and cannot be corrected to a normal distribution. This issue is discussed in detail in the supplementary materials, available at https://osf.io/nwv74/. For all of these reasons, our model is not perfect. But nevertheless, it makes remarkably good predictions for the f0 contours of Taiwan Mandarin disyllabic words with the RF tonal pattern, and therefore is useful in elucidating the factors that influence these contours, specifically enabling us to investigate the hitherto unrecognized contribution of meaning (see Box 1976, 1979 for discussion of the usefulness of statistical models).

The results presented so far provide strong evidence that word type is predictive of tonal realization, over and above the segment-related predictors established by previous studies, and that this is a robust effect, not attributable to overfitting. Our hypothesis, arising from the theoretical framework of the discriminative lexicon model (DLM; Baayen et al. 2019, Chuang & Baayen 2021, Heitmeier et al. 2026), is that the predictive power of word type arises not only from articulatory constraints, but also from a close association between word meaning and phonetic form, which enables the language learner or user to discriminate more efficiently between forms with different meanings. However, as discussed in Section 2.4, the effect of word type cannot be unequivocally attributed to semantics, since in addition to word meaning, word captures all of a word's form properties, not only those previously identified as affecting tonal realization. Furthermore, since word meaning varies with context, there is no one-to-one correspondence between word type and the meaning of a given token; word can encompass only a rather general approximation of what each token means. To address both of these issues, it is necessary to turn to prediction (ii) and the model with a factor smooth for sense as sole predictor added to the baseline. If word meaning is indeed predictive of tonal realization, then a model replacing the factor smooth for word by a factor smooth for sense should improve model fit even further.[24]

2.7. Results and discussion: sense

2.7.1. Evaluation of predictors

Figure 10 shows model fit improvement and concurvity for models based on the smaller data set that included at least fourteen tokens of each word sense. Even with this smaller data set, the overall pattern of results is very similar to that of the word-type analysis. Critically, however, sense appears to be a somewhat better predictor than word. For this data set, adding only a factor smooth for sense to the baseline model results in a fall of 6,443 AIC units, compared with a fall of 6,078 AIC units when only a factor smooth for word is added. The difference of 365 AIC units means that the sense model is 1.81e+79

[24] Note, however, that as discussed in Section 2.3, the senses that constitute the possible values of our sense variable discretize a much more subtle and interesting palette of shades of meanings.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

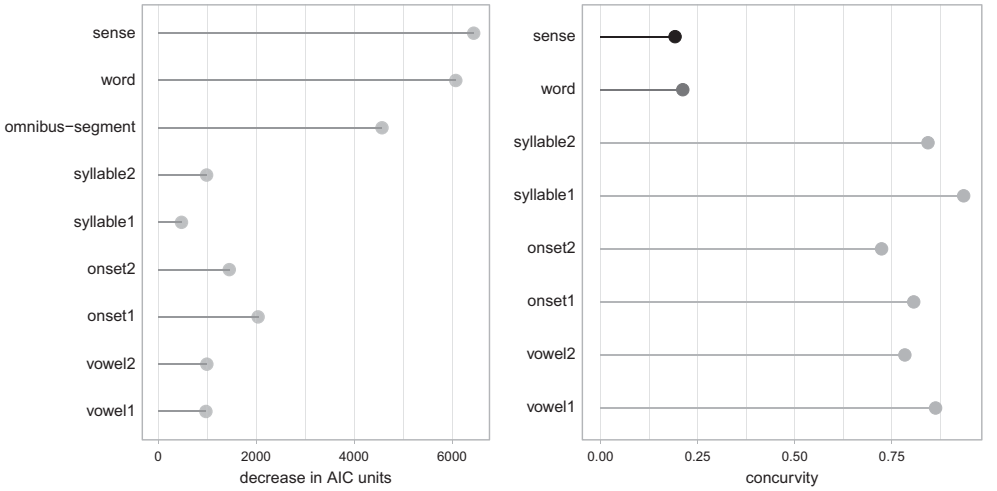


Figure 10. *The left-hand panel shows model fit improvement gauged by decrease in AIC units when a predictor (or set of predictors) is added to the baseline model for the sense analysis. The right-hand panel shows the concurvity score of individual predictors in three models using the smaller data set of 3,458 tokens: the omnibus-segment model with factor smooths for all segment-related control variables (light gray), the word-type model with a factor smooth for word predictor (dark gray), and the sense model with a factor smooth for sense (black).*

times more likely than the word model to explain the observed data. Furthermore, the effect of sense is also less confounded with the other predictors in the model (i.e. the baseline controls), as indicated by a smaller concurvity score.

Figure 11 presents the predicted tonal contours for different senses of three words: *bu2yao4* (left), *shi2zai4* (upper right), and *neng2gou4* (lower right). The word *bu2yao4* is a polysemous negation marker in Mandarin. The four senses that are found in our data set are 'prohibition', 'dissuasion', 'unneccesity', and 'to wish something to not happen' (s1 to s4, respectively). It can be seen that the different senses have clearly different tonal realizations. The panels on the right-hand side of Figure 11 present the predicted contours for the other two words, each of which has two senses in our data. For *shi2zai4*, tonal realizations vary greatly between the two senses ('truly' and 'indeed'), whereas the realizations of the two senses of *neng2gou4* ('being capable of' and 'enabling') are more alike and differ mainly with respect to the amplitude of the pitch inflection.

2.7.2. Cross-validation

As shown in Figure 12, for the 100 cross-validation runs, it turns out that the model with sense is not necessarily always more accurate than the model with word. The medians of the reduction in SSE for the word and the sense model are very similar, with the variance of the sense model being somewhat larger.

There are two reasons for the absence of greater prediction precision for models having access to sense instead of word. First, in the smaller data set used for these models, no fewer than thirty-five of the fifty-one word types are represented by only one sense. Any prediction advantage would therefore have to be contributed by just sixteen words. Second, for the majority of this subset of sixteen words, one sense accounts for most of the tokens. For tokens with these dominant senses, prediction is possible with greater precision. However, for tokens with less frequent senses, prediction is necessarily less precise. To see this, consider Figure 13, which presents predicted pitch contours and approximate confidence intervals for senses with many tokens (upper panels) and senses with few tokens (lower panels). Confidence bands are narrower for senses with many tokens. As a consequence, prediction for held-out tokens cannot be of the

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

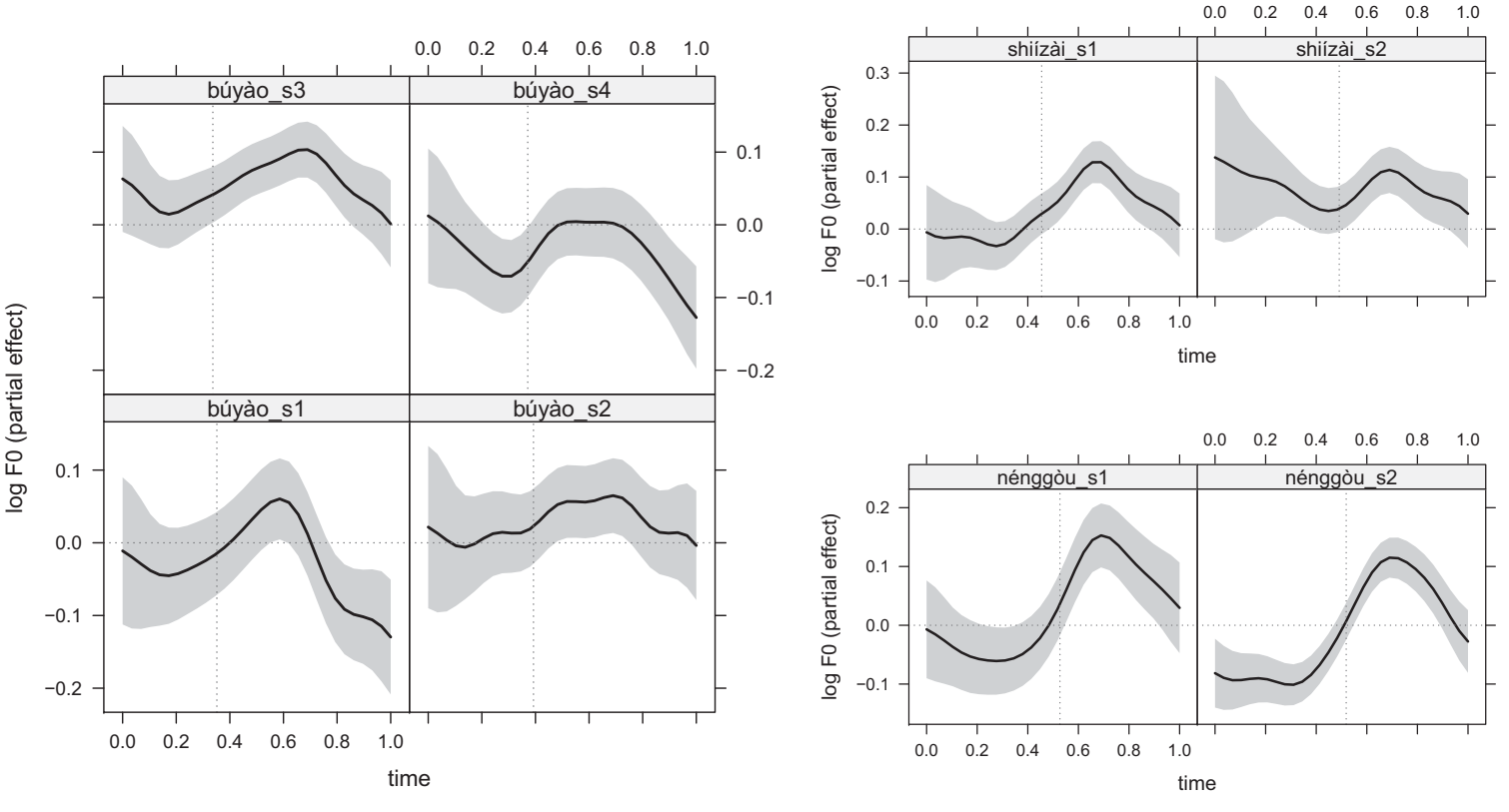


Figure 11. *Examples of the pitch contours predicted by the general smooth for time for female speakers, combined with the partial effects of the factor smooth for sense. The left-hand panel shows the fitted tonal contours for different senses of the word* bu2yao4*, a negation marker in Mandarin. The four senses are 'prohibition', 'dissuasion', 'unneccesity', and 'to wish something to not happen'. The upper right-hand panel shows the fitted tonal contours for the two senses of* shi2zai4*, meaning 'truly' and 'indeed', respectively. The lower right-hand panel plots the fitted contours for the two senses of* neng2gou4*: 'being capable of' and 'enabling'.*

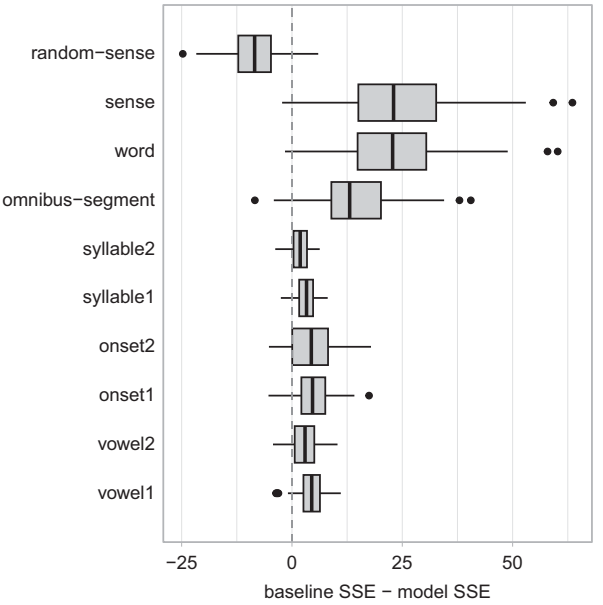


Figure 12. *Model accuracy under 100 runs of cross-validation for the sense analysis. The boxplots represent the distributions of reduction in SSE.*

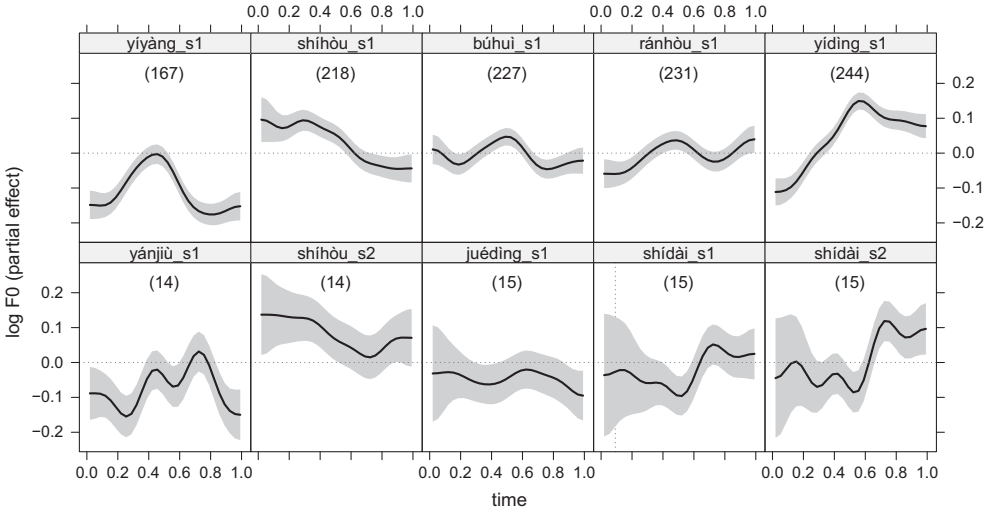


Figure 13. *Predicted pitch contours of the partial effects of the factor smooth for sense, for the five most frequent senses (upper row) and the five least frequent senses (lower row). Numbers in parentheses indicate the number of tokens in the data set for the different senses.*

same quality for senses with few tokens as compared to senses with many tokens. The overall improvement in model fit for the sense-based GAM results from the fact that the pitch contours of tokens with the dominant sense of each word can be better predicted once these tokens are separated from tokens of minority senses.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

2.8. Summary of Section 2

The results presented in this section have provided evidence in support of our first two predictions. First, word type is a stronger predictor of tonal realization than all of the previously established word-form-related predictors combined. Second, information about a word's meaning in context improves prediction of its tonal realization in that context. To the extent that sense labels provide more fine-grained meaning distinctions than word labels do, our results suggest that meaning plays a role in shaping the realization of tonal contours in Mandarin. In other words, in addition to the relevant segmental differences previously identified in the literature, differences in meaning also contribute.

Nevertheless, as discussed in Section 2.3, sense labels impose discrete categories on semantic variation that is actually much richer, more subtle, and more nebulous than can be captured by such inventories. In the computational models reported in Section 3, this problem did not arise, since our use of the DLM enabled us to replace relatively crude sense categories with token-specific semantic representations.

3. Understanding and producing item-specific f0 contours

So far we have shown that it is possible to identify meaning-specific modulations of the pitch contour for Mandarin words with the RF tonal pattern. The question therefore arises as to whether native speakers of Mandarin could in principle profit from these meaning-specific modulations. In other words, are the semantic components in words' pitch contours sufficiently informative that they could facilitate word comprehension for the listener? A related question is whether these subtle semantic modulations are learnable for a speaker, as opposed to arising mechanically each time a word is produced, as one might expect for purely articulatory effects. As outlined in Section 1, our third and fourth predictions, repeated here for convenience, anticipate an affirmative answer to both of these questions.

(iii) Given a pitch contour, the meaning of its carrier token can be predicted above chance level by a simple computational model with previous experience of that word type.
(iv) Assuming it has previous experience of the relevant word type, a simple computational model can produce an appropriate pitch contour for a given meaning.

In this section of the article, we explore these predictions with computational modeling, using the DLM (Baayen et al. 2019, Chuang & Baayen 2021, Heitmeier et al. 2026). If we can show that a simple computational model can learn to predict the meaning of a word token from its pitch contour and that pitch contours can be predicted from intended meaning, we have a proof of concept for the potential functionality of meaning-specific pitch realization in human lexical processing.

As described in Section 1, the DLM focuses on the relationship between words' forms and their meanings and allows for fine-grained alignments between low-level features of form and low-level features of meaning. Form-meaning relationships are captured by two networks: a comprehension network that maps word form onto word meaning, and a production network that maps word meaning onto word form. Recall that in the DLM theory of the mental lexicon, forms and meanings do not have representations in memory. Form representations represent ephemeral auditory or visual input, which dynamically generates a corresponding, equally ephemeral, meaning representation. Conversely, a meaning conceptualized by a speaker at a given point in time is dynamically transformed into ephemeral representations driving articulation. In line with this theory, the DLM generates forms and meanings on the fly on a token-by-token basis, making it possible to model the relationship between a given token's specific pitch contour and that token's context-specific meaning, and hence to account for correspondences between meaning and fine phonetic detail.[25]

[25] This is not possible in models of speech production and comprehension that rely on stored abstract representations to mediate between form and meaning, where every token of a given word type is assumed to be associated with the same stored representations (e.g. Cutler & Clifton 1999, Levelt et al. 1999).

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

In the DLM, both the form and the meaning of each word token are operationalized mathematically as high-dimensional numeric vectors. In order to test our predictions about the potential functionality of tonal modulations, the form vectors used in this study are based on the f0 contour of the relevant token. The meaning vectors that we use are context-specific, and hence also vary from token to token. Sections 3.1 and 3.2 describe how we obtained the vectors for meaning and pitch, respectively; Section 3.3 addresses the functionality of pitch in comprehension, and Section 3.4 does the same for production.

3.1. Representing meaning: contextualized embeddings

Embeddings are widely used numeric representations of words' meanings, developed from the distributional semantic insight that words with similar meanings tend to occur in similar contexts (Harris 1954, Firth 1968, Salton et al. 1975). Embeddings represent word meanings as real-valued high-dimensional vectors in a semantic space (Schütze 1992). They have been found to provide a plethora of novel insights in both psychology (Landauer & Dumais 1997, Bruni et al. 2014, Günther et al. 2019) and linguistics (Perek & Hilpert 2017, Nieder et al. 2023, Gahl & Baayen 2024) and are widely used in computer science (Bojanowski et al. 2017). First-generation word embeddings are static, type-level representations that model the meaning of a word type as a fixed point in semantic space, regardless of its usage in a given context. These representations therefore have difficulty distinguishing between multiple senses of a word (Pilehvar & Camacho-Collados 2020), and although various methods have been proposed to incorporate sense or context information into type-level embeddings (see e.g. Reisinger & Mooney 2010, Huang et al. 2012, Neelakantan et al. 2014, Iacobacci et al. 2015), most of these methods involve the use of sense-annotated corpora, which as far as we know are not available off the shelf for Mandarin and, in any case, have the disadvantage of discretizing more complex semantic variability. An alternative is to use **contextualized embeddings** (henceforth CEs). In contrast to static, type-level embeddings, which are based on word cooccurrences irrespective of order, CEs take into account the sequence of words in the immediate context of a target word. CEs therefore encode word meanings at the token level, and different tokens of the same word type will have different but similar context-specific embeddings.

To address the issue of words having context-specific meanings, this study used CEs produced for the tokens of our data by a pretrained unidirectional language model based on the GPT-2 architecture. The model, developed by CKIP, Academia Sinica, Taiwan,[26] was trained on a 4.3 billion character data set written with traditional Chinese characters. The model has 102 million parameters and encodes each character as a 768-dimensional vector. We presented the target words and their preceding contexts (consisting of all of the words that occur before the target in the current utterance, as well as all of the words in the immediately preceding utterance) to the GPT-2 model, and obtained two embeddings from the model, one for each character. Following standard machine-learning practice (e.g. Huang et al. 2021) we then averaged the two embeddings, so that every token in our data set received a 768-dimensional vector representing its context-specific meaning.

To visualize the semantic space of the CEs, we reduced the 768-dimensional semantic space to two dimensions using tSNE (van der Maaten & Hinton 2008). Figure 14 shows the resulting reduced two-dimensional plane, using convex hulls to highlight that tokens of different word types typically fall in distinct regions, while tokens of the same word type form clear clusters. Although perhaps unsurprising, this distribution confirms that, despite polysemy, the word types in our data do capture a general approximation of what each token means. It is also reassuring to see that, like static embeddings, the CEs can capture interword semantic relations. For instance, there is a cluster of school-related words in the lower left: *xue2xiao4* 'school', *yan2jiu4* 'research', and *xue2dao4* 'learn+resultative' (see Vulić et al. 2020 for similar results).

[26] We used ckiplab/gpt2-base-chinese, which is available on https://github.com/ckiplab/ckip-transformers.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

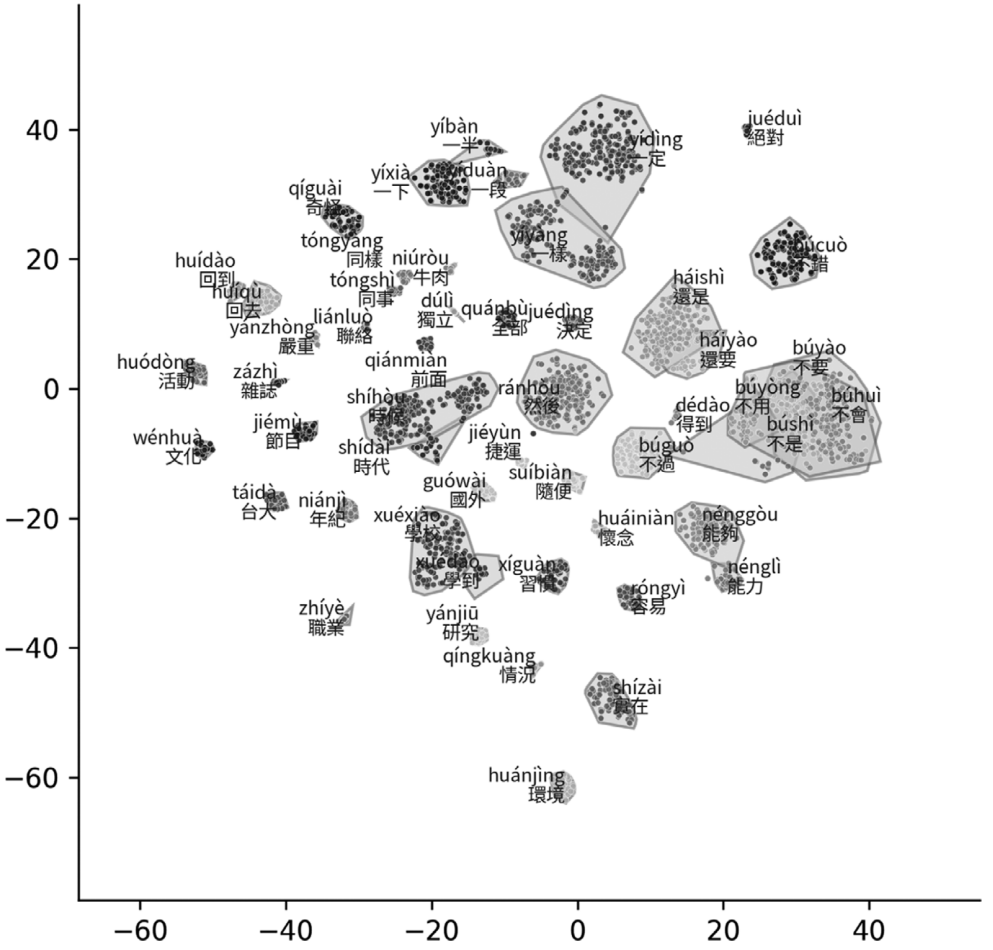


Figure 14. *Contextualized embeddings, obtained from a pretrained Chinese GPT-2 model, cluster by word type in the two-dimensional plane obtained with t-distributed stochastic neighbor embedding (van der Maaten & Hinton 2008). Convex hulls (gray polygons) show that the tokens of the different word types form well-localized and highly distinct clusters.*

3.2. Representing form: pitch vectors

For the DLM to implement mappings between form and meaning, every form-vector input to the model has to have the same number of dimensions as all of the others. However, because the tokens in our data vary in duration, our tokens also vary in the number of measurement points. This means that the raw measurements cannot be used to create the form vectors. The raw pitch contours also have the problem that there are gaps due to voicelessness. To overcome these problems, we used two of the GAMs described in Section 2 to obtain smoothed pitch contours from which we could extract a standard number of measurements. Although the sense GAM (Section 2.7) had a better fit to the data than the word GAM did (Section 2.6), the former unavoidably used a smaller data set than the latter. For our DLM models, we wanted to maximize the number of data points available; we therefore chose to use the word model and the corresponding omnibus-segment model. We generated two predicted pitch contours for each token, one using predictions from the word GAM, and the other using predictions from the omnibus-segment GAM. Each of these predicted contours was then used to generate f0 predictions at fifty equally spaced time points ranging between 0 and 1 for every individual token.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

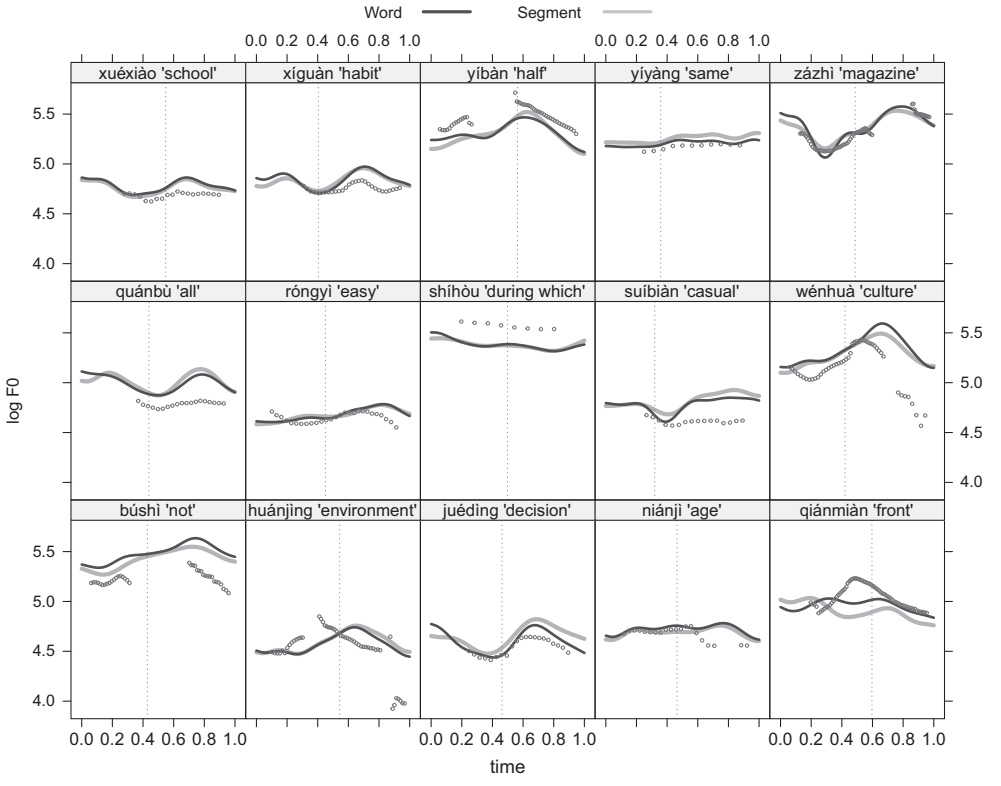


Figure 15. *One token randomly selected for a selection of words. The dots plot the observed pitch contour (raw data), and pitch vectors obtained from the word-type and the omnibus-segment models are represented by the dark gray and light gray curves, respectively. The vertical dotted lines indicate syllable boundaries.*

Both the word GAM and the omnibus-segment GAM include all of the speaker-related and context-related control variables described in Section 2.3. The only difference is that the former additionally includes word as the sole lexical predictor, while the latter includes six predictors specifying words' segmental properties. Examples of GAM-generated contours (from both the word and the omnibus-segment GAMs), together with the raw f0 values, are presented in Figure 15. As can be seen, the GAM-generated contours, though generally smoothing out the undulations in raw f0s, still largely capture the overall contour shape. Moreover, since the two GAMs provide similar but not identical predicted contours, it is possible to compare their performance in the DLM. If pitch contours generated from the word GAM provide superior fits to the respective semantic vectors compared to those generated from the omnibus-segment GAM, this will provide further evidence that the word variable is indeed contributing some meaning-related information.

A speaker's gender and individual characteristics such as vocal-tract anatomy, idiolect, and emotional state at the time of speaking all have strong effects both on their baseline pitch and on pitch range. Similarly, in both our word GAM and in the corresponding omnibus-segment GAM (Section 2.6), the intercepts are largely dependent on the speaker's gender and individual identity, and differences in amplitude are largely dependent on token duration, which we take to reflect both the speaker's idiolect and their emotional state at the time of speaking, among other things. On the semantic side, in normal spoken interaction between humans, a speaker's identity and emotional state not only contribute to the pitch contours they produce, but are also conceptually available to their interlocutors. In contrast, the CEs used as semantic representations in our DLM modeling (Section 3.1) are based entirely on written text and

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

therefore encode much less information about the speaker. To control for this discrepancy, we centered and scaled the predicted f0 values by token; that is, for each token in our data, and for each GAM, we calculated the mean and range of the fifty predicted f0 values, subtracted the mean from each predicted value, and divided the result by the range. This method of scaling (min-max normalization) ensures that the scaled data stays within a fixed range, between 0 and 1, so that every token contributes equally to the model fit, irrespective of its baseline pitch or amplitude; without scaling, tokens with a greater amplitude would be taken into account more than those with a lower amplitude. A consequence of the way we centered and scaled the pitch vectors is that our DLM production models generate predictions for the geometric shapes of the contours, but not for absolute pitch or amplitude.[27]

3.3. Modeling comprehension

3.3.1. Method

We used two different methods to map our pitch vectors onto our semantic vectors in a comprehension network. The first method involves a straightforward linear mapping using the **Linear Discriminative Learning** (henceforth LDL) engine of the DLM. This is equivalent to the standard linear mappings used in statistics for multivariate multiple regression (see e.g. Heitmeier et al. 2021, 2026, Gahl & Baayen 2024 for introductions). The second method (henceforth **ResLDL**) complements the linear mapping with an additional deep mapping, making it possible to accommodate nonlinear relations while keeping the model relatively interpretable. ResLDL augments an LDL mapping with a nonlinear deep network, which is given the task of capturing any systematicities that are left unexplained in the residuals of the linear network (hence the name ResLDL). Using both of these methods, and comparing the results, allowed us to shed light on the complexity of the relationship between our pitch vectors and our semantic vectors.[28]

We split our data into a training set (80%), a validation set (10%), and a test set (10%) in such a way that every word type was represented in all three sets of data and the number of tokens per word was proportional in all three sets. In other words, the test sets consisted entirely of novel tokens but no novel types, simulating the situation for human language use with previous experience. Both the LDL and the ResLDL mappings were trained on the training data and tested on the test data. In accordance with standard machine-learning practice, the validation set was used to fine-tune the hyperparameters in the ResLDL model before testing. This was not necessary for the LDL model, since there are no hyperparameters in LDL. To ensure that our results were not specific to a particular data split, we repeated the entire modeling procedure thirty times using repeated training/test splits, that is, Monte Carlo cross-validation (Zhang 1993, Kuhn & Johnson 2013). The repeated splits followed the same proportions described above, generating thirty accuracy scores for each combination of pitch type (omnibus-segment or word f0 smooths) and network (LDL or ResLDL).

We evaluated the accuracy of model predictions as follows. For each pitch vector in the test set, we obtained a corresponding predicted semantic vector and identified its closest neighbor among the actual CEs of the tokens in our data. If this nearest neighbor belonged to any token of the same word type as the target token, the predicted semantic vector was assessed as correct, and otherwise as incorrect. This measure of success was chosen for both computational and conceptual reasons, as detailed in the following two paragraphs.

Although one might expect that a predicted semantic vector would ideally be closest in semantic space to the CE of the held-out token in question, this is computationally unrealistic. The CEs in our models are conditioned on the preceding context of a given token and are uninformed about the following context. The pitch contours, in contrast, are shaped in part by the tone on the following word and the probability of the word given the next word. Thus, there is information in the pitch contours that is absent in the CEs,

[27] Geometric shape can be defined as 'all the geometrical information that remains when location, scale, and rotational effects are filtered out from an object' (adapted from Kendall 1977:428).

[28] Technical details of the LDL and ResLDL implementation can be found in the online supplementary material at https://osf.io/nwv74/.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

making it computationally infeasible to predict token-specific vectors. Furthermore, both the pitch contours and the CEs have measurement error. Similar to the way that a linear regression line predicts the mean value of a dependent variable for a given value of an independent variable, but not the individual data points used to generate the line, here we can predict at the level of types, but not at the level of individual tokens.

From a cognitive perspective, it is worth noting that listeners cannot arrive at exactly the same conceptualization as the speaker, as listeners and speakers have different experiences with the language and different life histories. For example, a listener who hears 'Do you fancy a coffee?' may conceptualize a cappuccino, even if the speaker was envisioning an espresso. Fortunately, provided both interlocutors arrive at similar-enough meanings, communication can proceed unhindered. In addition to being computationally feasible, using same word type as a criterion for success therefore makes sense in terms of human performance levels.

3.3.2. Results

Figure 16 presents the mean comprehension accuracies for the training data (left) and the test data (right). The individual barplots show accuracies for LDL (left two bars) and ResLDL (right two bars), for pitch contours based on segment-aware GAMs (black) and on word-aware GAMs (gray). As mentioned above, a given prediction is considered correct when the closest neighbor of the predicted CE is of the same word type as the target. For LDL, accuracy hovers around 30% for both training and test data, whereas for ResLDL accuracy is higher, over 60% and 50% for training and test data, respectively. These results are surprisingly good, given that the models are requested to predict semantic vectors on the basis of pitch information only, notwithstanding the fact that pitch contains implicit information about phonological segments (cf. Section 2.6). For comparison, across our whole data set the theoretical probability of a pitch vector and CE belonging to the same word type by chance is approximately 0.038. Similarly, baseline accuracies obtained by evaluating on a data set with randomly permuted word labels were 3.7% for the training set and 3.5% for the test set. This allows us to conclude that the classification accuracies of our models are far from trivial. On the contrary, even the least successful model achieves accuracies that are a whole order of magnitude greater than would be expected by chance.

The higher accuracies of the ResLDL model compared to the LDL model indicate that mappings from pitch contours to CEs have significant nonlinear components. This nonlinear mapping may be required

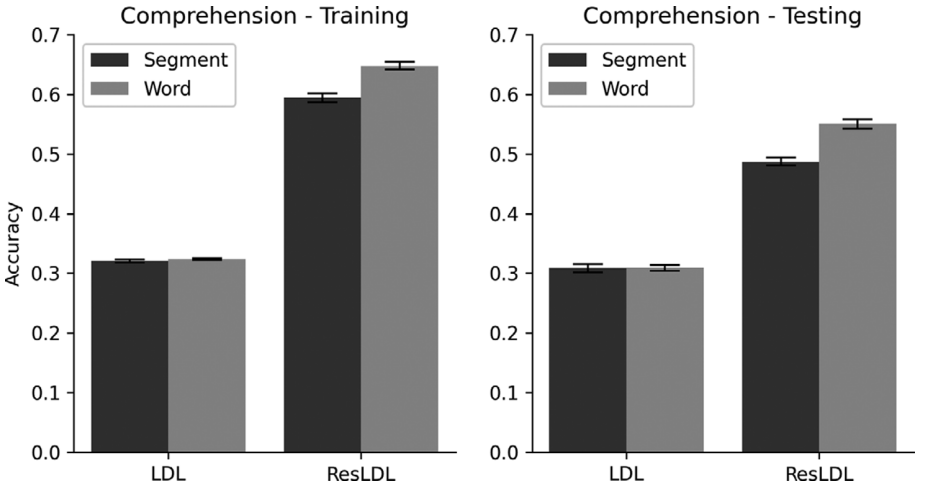


Figure 16. *Mean comprehension accuracies for training data (left) and test data (right) for LDL and ResLDL mappings from omnibus-segment (black) and word (gray) pitch vectors. Mean accuracy is obtained from thirty stratified random training and testing splits, each trained and evaluated independently. Error bars indicate double the standard error.*

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

because we are mapping from fifty-dimensional pitch contours to 768-dimensional CEs. As it is impossible to map a lower-dimensional space into a higher-dimensional space with a linear mapping without losing information,[29] the greater accuracy of ResLDL is unsurprising. Nevertheless, it is remarkable that the linear mappings show very similar performance on training and test data, suggesting that there is a strong linear component to predicting meaning from tonal contours.

A comparison of results from the two types of pitch vectors shows that meaning prediction with ResLDL is more accurate when the pitch contour smooths are generated using the word GAM than using the omnibus-segment GAM. This indicates that the factor smooth for word not only contributes to a better model fit in the GAM, but also produces predicted pitch contours that are better aligned with words' meanings, albeit in a nonlinear way. Of course, word encodes all of the information about words' segmental makeup that is given to the omnibus-segment model, so the resulting contours do contain this information. But the superior performance of the word-based contours shows that tonal realizations that include all information associated with word type have the potential to help listeners to identify words' meanings even more accurately than contours with just segmental information.

3.4. Modeling production

3.4.1. Method

We have seen that the pitch contours of Mandarin disyllabic words contain substantial information about word meaning. It is remarkable that a DLM comprehension model can achieve a test accuracy of over 50% when modeling with word-aware pitch contours and ResLDL. We now turn to production, addressing the question of whether a token's pitch contour can be predicted with reasonable accuracy from its CE. If so, this would support our hypothesis that speakers can in principle learn to produce meaning-specific tonal contours.

Before going into further detail, we note that this task is considerably more difficult than the task presented to the word GAM model in Section 2. The GAM model was asked to predict pitch contours from word labels and was oblivious to variation in meaning between tokens of a given word type. In the models reported below, however, the LDL and ResLDL mappings are confronted with semantic vectors that are different from token to token. The question is whether the similarities between the CEs of tokens belonging to the same word are sufficiently consistent for the LDL and ResLDL mappings to predict appropriately similar pitch contours.

Model set-up was the same as for comprehension, except that to model production the input consisted of CEs and the output consisted of pitch vectors. We again conducted the modeling procedure thirty times. For each CE in the test set, we obtained a corresponding predicted pitch vector and identified its closest neighbor among the actual (GAM-generated) contours of the tokens in our data. If this nearest neighbor belonged to any token of the same word type as the target token, the predicted pitch vector was assessed as correct, and otherwise as incorrect.

We complemented the quantitative evaluation with a qualitative analysis of the pitch contours predicted by the model for individual word types. To do this, we calculated the centroid of the CEs for all tokens of a given word, and used this centroid vector to generate a predicted pitch contour from the production network with LDL mappings and word-based pitch vectors. For each of the words presented in Figure 5 above, we then assessed the quality of this LDL-predicted contour by visually comparing it with the contour produced by averaging the actual (GAM-generated) pitch vectors used to train the model, for all tokens of the word in question.

3.4.2. Results

Mean production accuracies (over thirty repetitions) for the token-based evaluation are presented in Figure 17. For training data (left), accuracies are between 40% and 50%. The accuracies for the test data

[29] To see this, consider points in a cube and their projection onto a plane in that cube. From that projection, which is two dimensional, the original locations of the points in the three-dimensional cube cannot be fully reconstructed.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

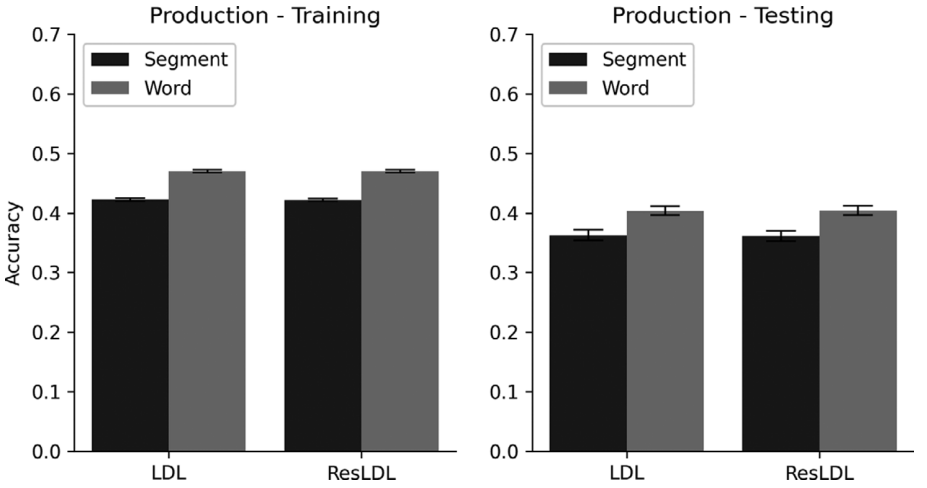


Figure 17. *Mean production accuracies for training data (left) and test data (right) for LDL and ResLDL mappings from omnibus-segment (black) and word-type (gray) pitch vectors. Mean accuracy is obtained from thirty stratified random training and testing splits, each trained and evaluated independently. Error bars indicate double the standard error.*

are only slightly lower, hovering around 40%. The probability of a CE and pitch vector belonging to the same word type by chance is the same as for the comprehension models, namely 0.038. Permutation baselines are again 3.7% for training and 3.5% for testing. In other words, like the comprehension models, the production models have accuracies an order of magnitude greater than would be expected due to chance. However, in contrast to the comprehension results, production accuracies are remarkably similar for LDL and ResLDL. Apparently, linear mappings suffice when predicting low-dimensional pitch contours from high-dimensional CEs and succeed in capturing the regularities in the meaning-to-form mappings. Possibly, predicting pitch from semantics is a cognitively more natural task than predicting semantics from just pitch on its own, and hence requires less powerful mappings. Finally, as for comprehension mappings, predicting pitch contours from CEs is more successful when pitch contours are generated with word-based GAMs, compared to segment-based GAMs.

The results of the qualitative analysis are shown in Figure 18. The LDL-predicted contours are shown in dark gray, and the by-type averages of the contours used to train the model are shown in light gray. A comparison of these two contours for any given word reveals remarkable similarity, indicating that the LDL production model generates high-quality predictions for the shapes of the pitch contours. It is also striking that, in shape, these contours closely resemble the contours in Figure 5, reproduced for convenience as the black contours in Figure 18.[30] Recall from Section 2.6 that this third set of contours was produced by combining the partial effect smooth for each word type with the general smooth for time for female speakers. The similarity therefore suggests that the word-specific pitch contours isolated by our word GAM can be understood as pitch contours that correspond to the centroids of word's contextualized embeddings. From this, we draw the conclusion that there is considerable isomorphy between the space of token-specific pitch contours and the semantic space of token-specific embeddings.

In addition to by-word centroids, we also calculated the centroid of the CEs for all tokens of all types in our data set. The pitch contour predicted from this overall centroid is very similar to the pitch contour in the right-hand panel of Figure 1 above. We therefore infer that the centroid of the embeddings of all tokens can be interpreted as the 'meaning' of the unmodulated rise-fall pitch contour. The ten tokens that are closest to this centroid belong to the word types *bu2guo4* 'but, however', *ran2hou4* 'and then', and

[30] Note that the black contours in Figure 18 are identical to the contours in Figure 5, except that the amplitudes are different due to the way we centered and scaled the contours in Figure 18.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

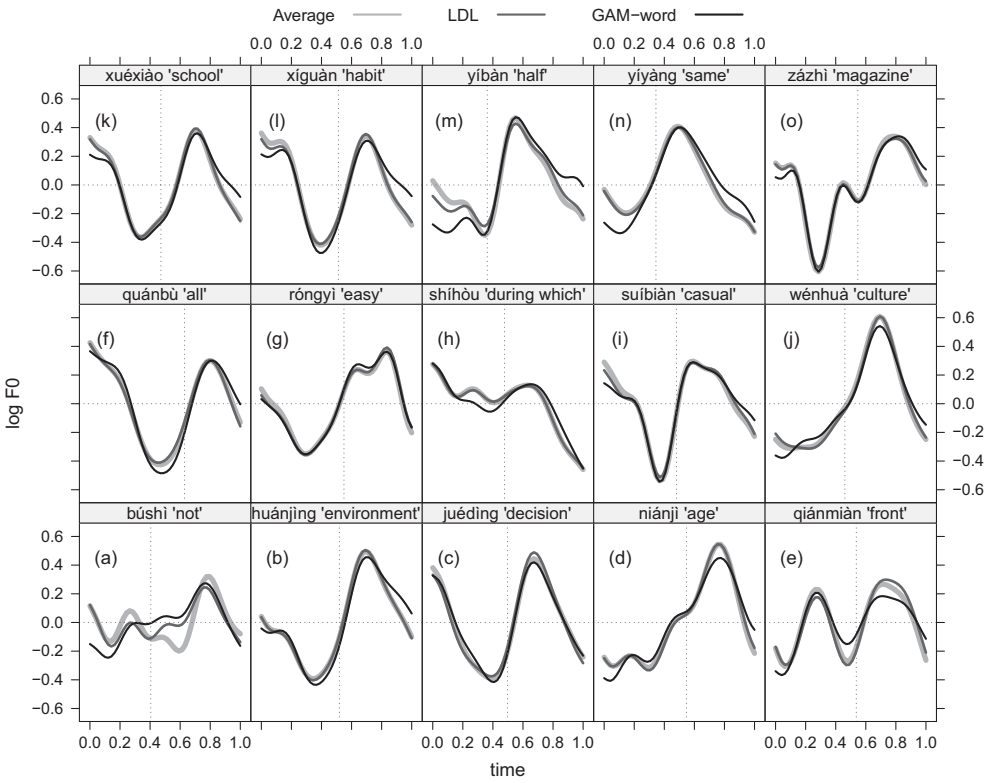


Figure 18. *Pitch contours for the sample of fifteen word types introduced in Figure 5. The light gray lines represent the average of the pitch vectors generated by the word-type GAM across all tokens of that type (i.e. the average of the contours used to train LDL). The dark gray lines represent the predictions generated by LDL. These LDL contours were predicted from 'centroid' word meaning, obtained by averaging the CEs of all tokens of the same type. The black lines represent the word-specific contours predicted by the word GAM as presented in Figure 5 and reproduced here after centering and scaling. That is, these black lines show the pure effect of word on the pitch contour, irrespective of other predictors. The vertical dotted lines in the panels indicate the average (word-specific) syllable boundary.*

shi2hou4 'during which', which suggests that these words are the most typical carriers of the RF pitch contour in the current data set.

4. General discussion

This study investigated variation in the f0 contours of disyllabic words with the rise-fall (RF) tonal pattern in Taiwan Mandarin. The central hypothesis of our study is that Taiwan Mandarin disyllabic word tokens have pitch contours that are in part driven by their meanings. In standard analyses of tone in Mandarin, the rising tone of the first syllable and the falling tone of the second syllable of RF words are inherited from the single-syllable constituents and are taken to be basic, underlying tones. Deviations from these tones are explained by appealing to articulatory and prosodic constraints governing how tones can be realized. Our hypothesis adds word meaning as a missing player in the articulatory arena by arguing that meaning co-determines the realization of Mandarin tones.

Our core hypothesis generates four predictions. The first prediction is that word type will be a stronger predictor of tonal realization than all of the previously established word-form-related predictors combined. This prediction follows from the hypothesis because word type includes information about

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

meaning in addition to information about form. Using generalized additive models (GAMs), we were able to show that word type is indeed a more powerful predictor of tonal realization than a wide range of words' form properties considered jointly. We not only established that the GAM with a factor smooth for word type provided a substantially improved model fit, but also demonstrated that the word-informed GAM provided more accurate predictions for the f0 contours of held-out data. We concluded that individual word types have specific properties—over and above their segmental makeup—that modulate the general, sine-wave shaped f0 contour characteristic of words described as having a rise-fall tonal pattern. These specific properties, we conjectured, are semantic in nature.

The second prediction is that information about a word's meaning in context will improve prediction of its tonal realization. If words with the very same segments and canonical tones, but different meanings, have distinct tonal realizations, this provides evidence for the possibility that there is a semantic component to the realization of tone in Taiwan Mandarin disyllabic words. Again using generalized additive modeling, we were able to show that adding information about meaning in context, that is, sense, did lead to significant improvement in model fit. These results provide evidence for the possibility that Mandarin disyllabic word tokens indeed have tonal realizations that are partially determined by their semantics.[31]

The third prediction is that given a pitch contour, the meaning of its carrier token can be predicted above chance level by a simple computational model with previous experience of that word type. To test this, we used the framework of the discriminative lexicon model. Given the difficult task of predicting words' high-dimensional semantic embeddings from low-dimensional pitch contours, the DLM comprehension model that we implemented performed on held-out data with an accuracy of over 50%, compared with a random baseline of 3.5%. The tonal contours of word tokens turned out to be far more revealing about their meanings than anything we thought might be possible when we started this investigation.

This finding has two important implications. First, for the model to learn to predict meanings from pitch contours, pitch contours must contain information that aligns with aspects of word meaning. This suggests that human speakers of Mandarin could also potentially make use of this information to optimize speech comprehension. Second, meaning-specific pitch contours might be related to the extensive homophony in Mandarin. According to the Chinese Lexical Database (Sun et al. 2018), about 90% of monosyllabic Mandarin words have at least one homophonous counterpart. From a functional perspective, the presence of meaning-specific pitch modulations may therefore compensate for the lack of semantic discriminability afforded by segmental makeup and syllable structure (see e.g. Sampson 2015, 2019 for a discussion of theoretical implications). However, given that the present study focuses on disyllabic words, which exhibit much less homophony than monosyllabic words, it remains an empirical question whether word-specific pitch contours for disyllabic words have the same functional load as contours for monosyllabic words would have. Ongoing research (Jin et al. 2025) strongly suggests that word-specific pitch contours are also present for monomorphemic words.

The fourth and final prediction that follows from our central hypothesis is that, assuming it has previous experience of the relevant word type, a simple computational model can produce an appropriate pitch contour for a given meaning. We again tested this prediction using the DLM. A network trained to match context-specific semantic embeddings to token-specific pitch contours performed far above a random baseline, with accuracies ranging from 40% to 45% on training data, and 35% to 40% on testing data. Given that the computational models were forced to predict pitch across tokens produced by many different speakers, without any information about the segmental makeup or syllable structure of the words, this is a remarkable result that provides strong support for human speakers in principle being able to learn to produce meaning-specific pitch contours. At this point, however, a word of caution is appropriate. Our DLM models were given the task of predicting the geometric shape of pitch contours

[31] We note that similar findings have been obtained in other studies as well. Lu et al. (2025b), for example, investigated all twenty tonal patterns of disyllabic Mandarin words and reported strong effects for word type and word sense. Jin et al. (2025), looking into monosyllabic words in Mandarin, also found word-specific tonal realizations.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

and did not address token- and word-specific differences in pitch height and amplitude. The modeling of the full pitch contours, including height and amplitude, is left for future investigation.

Although the four predictions that follow from our central hypothesis are empirically well supported, this does not necessarily imply that our hypothesis is correct. We cannot rule out that the importance of meaning in our models might actually be due to factors that we did not take into account in our analyses. For instance, the effects of prosody, pragmatics, syntax, and emotion could, in principle, conspire to yield effects that would seem to imply token-specific semantic effects. Measures of surprisal and informativity other than the forward and backward probabilities that we included as control variables may also be informative (Tang & Shaw 2021). In addition, it could be argued that in the present study, which is based on spontaneous conversational speech, the consequences of contraction and reduction (Ernestus 2000, Johnson 2004, Tseng 2005a) are not controlled for. And indeed, we agree that all of these factors are worth further investigation. However, it seems unlikely to us that any of these factors will turn out to explain away completely the effect of meaning. Our analyses are based on multiple tokens of each word type, which vary with respect to their syntactic position, their pragmatic function, the amount of segmental reduction, and their emotional valence. The pitch modulations estimated with the help of the GAM models are statistical generalizations across all of this variation that is present in our data. It is unlikely that the factors we were not able to control for will be distributed across our tokens in such an unbalanced way that they would be able to explain away our semantic effects. But even were it to turn out that tonal variation is completely predictable from factors other than semantics, our key point is still valid: our DLM models show that this word-specific variation is fine-tuned with words' meanings as represented by contextualized embeddings.

In our DLM comprehension model, the mapping from a pitch contour to its context-specific embedding is to a large extent linear; in the production model, the reverse mapping is completely linear. These facts indicate that there is considerable isomorphy between the form space of Mandarin word tokens' pitch contours and the semantic space of Mandarin words' context-specific meanings. In other words, form and meaning mirror each other to a much greater extent than is often assumed, especially in frameworks that take as axiomatic that language has a 'dual articulation' (Martinet 1965) that allocates form and meaning to two unrelated, orthogonal, components of the grammar. Furthermore, the existence of this isomorphy means that human language users could in principle exploit the associations between form and meaning to optimize comprehension and production.

The question then arises as to whether listeners actually do make use of the distributional-statistical information that is in the speech signal of Mandarin RF words: we think there is a strong possibility that they do so. The pertinent information is present in the speech signal, so speakers are producing tonal realizations that align with word meaning. But this isomorphy cannot be reduced to the consequences of biomechanical constraints on the speech production process. Listeners must be learning the distributional statistics of tone and meaning from the speech to which they are exposed. We hasten to note that the learning of the systematicities between pitch and semantics is in all likelihood a completely subliminal process. It is not necessary for learners to be aware of the subtle modulations of pitch contours in relation to equally subtle nuances in meaning. In our conception of the learning process, successful understanding, token by token, will drive low-level learning in the lexical networks, without conscious reflection and effort being required.[32]

Since our central hypothesis concerns word meaning, it was essential at the outset of this article to clarify our theoretical position apropos of semantics in general and word meaning in particular. As outlined in Section 1, we adopt the framework of contextualism; that is, in contrast to theories that conceptualize word meanings as abstract context-independent symbols, we take the view that what a word means varies with its context, a view reflected in our use of contextualized embeddings to represent word meanings in our computational models. This conceptualization of word meanings as context-dependent high-dimensional

[32] For token-by-token incremental learning in lexical decision mega experiments, see Heitmeier et al. 2023, and for continuous recalibration in vision, see Marsolek 2008.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

vectors has many philosophical, cognitive, and empirical advantages, discussed in detail by Landauer and Dumais (1997) and by Günther et al. (2019), who address common misunderstandings.

In the present study, the use of contextualized embeddings has offered novel insights into the production of tone in Taiwan Mandarin, but this is just one illustration of their application in elucidating linguistic patterns that can be missed if semantics is taken to be independent of context. For example, in the domain of auditory perception, de Varda and Marelli (2025) used embeddings to shed new light on the auditory iconicity characterizing spoken English. In the domain of inflectional morphology, Shafaei-Bajestan et al. (2024) also used embeddings to show that in the semantic space of English, the shift from singular to plural varies systematically with semantic class. In English, these differences in plural semantics are not grammaticalized, but languages such as Swahili and Kiowa have nominal inflection classes that have different plural exponents depending on semantic class. Embeddings bring to light striking similarities between English and these languages that are invisible to theories assuming that word meanings are invariant abstract symbols.

Although we are advocating a contextualist view of semantics, it is also possible in the DLM to represent word meanings as context-independent abstract symbols using one-hot encoded binary vectors in a semantic space that has as many dimensions as there are meanings; that is, each word is represented by '1' in the dimension corresponding to the relevant meaning and '0' in every other dimension. In this very high dimensional space, all word meanings are completely unrelated and orthogonal to each other. These simple representations can be surprisingly effective (see e.g. the naive discriminative learning model; Baayen et al. 2011), but we are not aware of any way we could use them to predict pitch contours in conversational Taiwan Mandarin. Furthermore, at a conceptual level, it is difficult to see how theories that adopt as foundational the axiom that meanings are abstract context-independent symbols would ever be able to offer the precise predictions enabled by embeddings, and hence account for the findings of the present study.

We have so far not discussed the consequences of our findings for phonological theories of Mandarin tone. Standard theories assume that Mandarin disyllabic words inherit the lexical tones of their constituent syllables, and that these tones underlie the pitch contours observed in spoken tokens. Variation in the phonetic realization of tones is attributed to the voluntary and involuntary processes described in Section 1, including biomechanical constraints on their production. In contrast, in the DLM models presented in this study, tonal realizations of disyllabic words are generated from their meanings. In the following paragraphs, we consider how such a model can account for three important aspects of Mandarin pronunciation: first, the tonal patterns that have been well documented for laboratory speech (see e.g. Xu 1997); second, tone sandhi, such as the tonal dissimilation of T3–T3 to T2–T3; and third, the biomechanical factors known to co-determine the realization of pitch.

The issue of tonal patterns was addressed in detail by Lu et al. (2025b), who used the same methodology as developed in the present study to investigate the tonal realization in conversational Taiwan Mandarin of the twenty tonal patterns possible for disyllabic words. They observed a clear, albeit small, partial effect of tonal pattern, alongside a much larger effect of word. As in the present study, the T2–T4 pattern was realized as a fall-rise-fall (see e.g. Figure 3), while the other tonal patterns likewise exhibited distinct phonetic signatures. Lu et al. (2025b) showed that the phonetic signature of any given pattern could be predicted from the centroid of the CEs of the tokens representing that pattern, just as the fall-rise-fall pattern is predicted by the centroid of the CEs of the T2–T3 tokens in our study (Section 3.4). In other words, the characteristic tonal patterns in the form space of Taiwan Mandarin correspond to patterns in semantic space; in the DLM, this isomorphy is dynamically encoded in the connection weights of the production and comprehension networks, and as a consequence, there is no need for the DLM to store tonal patterns in the form of discrete representations. In fact, as explained in Section 1, the DLM does not store representations of any kind; correspondingly, in our theory of the mental lexicon there are no stored representations, neither abstract underlying representations nor exemplars.

We turn now to the issue of tone sandhi. Lu et al. (2025b) showed that the centroids of the T2–T3 and T3–T3 tonal patterns in semantic space are very similar; in the framework of the DLM, it is therefore expected that the corresponding pitch signatures of these two tonal patterns will also be very similar

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

(in fact, for Taiwan Mandarin, identical; see Lu et al. 2025a). By contrast, tone sandhi in Mandarin Chinese is sometimes seen as evidence of underlying tonal representations, especially when it occurs productively in novel words, for example in the **wug test** (Zhang & Lai 2010, Zhang et al. 2015). While the wug test has been modeled in the framework of the DLM by Heitmeier et al. (2021, 2026), it remains an empirical question whether their approach can be extended to Mandarin tone; wug tests with phonotactically legal novel words, such as those used by Zhang and Lai (2010) and Zhang et al. (2015), are likely to have a strong metalinguistic component and may therefore be inaccessible to the DLM, which is a model of subliminal, automatized, lexical processing. However, we note that the wug test results cited as evidence for underlying tones are by no means universally replicated. For example, Hsieh (1976) and Wang (1972) found for Taiwanese Southern Min[33] that the success rate of the wug test is overall low, even for native speakers, and varies substantially across different sandhi rules.[34]

Concerning the biomechanical processes that constrain the realization of pitch contours, our hypothesis is that the collocational constraints that underlie CEs have their mirror image in biomechanical constraints. To test this hypothesis for internal tone sandhi, the general approach laid out in the present study could be pursued, but we anticipate that for real progress, improved speaker-specific CEs would be necessary, rather than the artificial intelligence CEs that we have had to rely on. To capture external tone sandhi, it will be essential to work with small utterances instead of two-syllable words.

The results reported in the present study have important consequences for the teaching of Mandarin as a second language. If words have individual pitch signatures, then presenting second language learners with the tones indicated by Pinyin transcriptions will be highly confusing and counterproductive. For instance, a learner presented with the Pinyin of *xuéxiào* 'school', but hearing a fall-rise-fall and noticing the initial descending pitch, will be completely confused by the discrepancy between what they are hearing and what (according to the Pinyin) they should be hearing. In fact, for many tokens, the tone representation in Pinyin will stand in the way of learning how Mandarin speakers actually realize tones, which may be why many language-learning apps such as Duolingo do not provide feedback on learners' tone production. When error feedback to the learner is itself error-ridden, not much progress can be expected. It is an empirical question whether presenting word-specific tonal realizations would truly improve tone learning for L2 learners; however, we believe that this approach is promising and should be able to provide new insights to inform pedagogical methods.

To conclude, we have provided a range of observations that are consistent with the possibility that the details of how tones are realized in Taiwan Mandarin disyllabic words is partially determined by meaning in context. If our interpretation of these observations is on the right track, semantics is an important missing player in current phonetic studies of f0 modulation in Mandarin. We believe our empirical findings are sufficiently strong to open up new lines of research on the realization of pitch in tone languages. We also hope that our findings will contribute to an improved understanding of why deep learning speech-processing systems are so remarkably effective and now constitute the state of the art in natural language processing. Our hypothesis is that these systems can pick up systematicities between form and meaning that are not open to human introspection, but that are visible to GAMs and computational models. Crucially, we hypothesize that these systematicities not only are exploited by computational modeling algorithms, but are also essential, albeit subliminally, for optimizing human lexical processing in comprehension and production.

Data availability statement. The data sets and supplementary materials for this study are available at https://osf.io/nwv74/.

Acknowledgments. The authors are indebted to Dr. Matteo Fasiolo, University of Bristol, for his statistical advice on the application of GAMs to pitch contours. Many thanks go to Jingwen Li for her effort devoted to data collection and analyses at the beginning of the project. We also thank three anonymous referees and the associate editor, Morgan Sonderegger, for their extensive and constructive feedback on this study. [Full editorial history: Received 04 May 2024; revision invited 03 October 2024;

[33] Taiwanese Southern Min, also known as Taiwanese Hokkien or simply Taiwanese, is a tone language genealogically related to the Min Chinese language.

[34] For discussion of the limitations of wug tasks, see also Nieder et al. 2023.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

revision received 28 March 2025; revision invited 09 July 2025; revision received 16 August 2025; accepted pending revisions 05 October 2025; revision received 10 October 2025; accepted 19 November 2025].

Funding disclosure statement. This research was funded by the European Research Council, grant SUBLIMINAL (#101054902) awarded to Harald Baayen, and partially supported by the Yushan Fellow Program by the Ministry of Education of Taiwan awarded to Yu-Ying Chuang.

Conflict of interest. The authors declare no conflicts of interest.

Ethics statement. The authors were given access to the Taiwan Mandarin Spontaneous Speech Corpus (Fon 2004), under the condition that the corpus is not made publicly available, in order to protect the privacy of the speakers.

References

Akaike, Hirotogu. 1998. Information theory and an extension of the maximum likelihood principle. *Selected papers of Hirotugu Akaike*, ed. by Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa, 199–213. New York: Springer. https://doi.org/10.1007/978-1-4612-1694-0_15.

Baayen, R. Harald; Yu-Ying Chuang; Elnaz Shafaei-Bajestan; and James P. Blevins. 2019. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity* 2019:4895891. https://doi.org/10.1155/2019/4895891.

Baayen, R. Harald; Matteo Fasiolo; Simon Wood; and Yu-Ying Chuang. 2022. A note on the modeling of the effects of experimental time in psycholinguistic experiments. *The Mental Lexicon* 17.178–212. https://doi.org/10.1075/ml.21012.baa.

Baayen, R. Harald; Petar Milin; Dusica Filipović Đurđević; Peter Hendrix; and Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118.438–82. https://doi.org/10.1037/a0023851.

Bell, Alan; Daniel Jurafsky; Eric Fosler-Lussier; Cynthia Girand; Michelle Gregory; and Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America* 113.1001–24. https://doi.org/10.1121/1.1534836.

Bi, Yifei; Yiya Chen; and Niels O. Schiller. 2015. The effect of word frequency and neighbourhood density on tone merge. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0636.pdf.

Boersma, Paul, and David Weenink. 2019. Praat: Doing phonetics by computer [computer program]. http://www.praat.org/.

Bojanowski, Piotr; Edouard Grave; Armand Joulin; and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5.135–46. https://doi.org/10.1162/tacl_a_00051.

Box, George E. P. 1976. Science and statistics. *Journal of the American Statistical Association* 71.791–99. https://doi.org/10.1080/01621459.1976.10480949.

Box, George E. P. 1979. Robustness in the strategy of scientific model building. *Robustness in statistics*, ed. by Robert L. Launer and Graham N. Wilkinson, 201–36. London: Academic Press. https://doi.org/10.1016/B978-0-12-438150-6.50018-2.

Bruni, Elia; Nam-Khanh Tran; and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 49.1–47. https://doi.org/10.1613/jair.4135.

Chao, Yuen Ren. 1968. *A grammar of spoken Chinese*. Berkeley: University of California Press.

Chen, Yiya. 2010. Post-focus F_0 compression—Now you see it, now you don't. *Journal of Phonetics* 38.517–25. https://doi.org/10.1016/j.wocn.2010.06.004.

Cheng, Chierh, and Yi Xu. 2015. Mechanism of disyllabic tonal reduction in Taiwan Mandarin. *Language and Speech* 58.281–314. https://doi.org/10.1177/0023830914543286.

Chuang, Yu-Ying, and R. Harald Baayen. 2021. Discriminative learning and the lexicon: NDL and LDL. *Oxford research encyclopedia of linguistics*. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.375.

Chuang, Yu-Ying; Yi-Hsuan Huang; and Janice Fon. 2007. The effect of incredulity and particle on the intonation of yes/no questions in Taiwan Mandarin. *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, 1261–64. https://www.icphs2007.de/.

Chuang, Yu-Ying; Mihi Kang; Xuefeng Luo; and R. Harald Baayen. 2023. Vector space morphology with linear discriminative learning. *Linguistic morphology in the mind and brain*, ed. by Davide Crepaldi, 167–83. London: Routledge. https://doi.org/10.4324/9781003159759-12.

Chung, Karen Steffen. 2006. Contraction and backgrounding in Taiwan Mandarin. *Concentric: Studies in Linguistics* 32.69–88.

Cutler, Anne, and Charles Clifton, Jr. 1999. Comprehending spoken language: A blueprint of the listener. *The neurocognition of language*, ed. by Colin M. Brown and Peter Hagoort, 123–66. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198507932.003.0005.

de Varda, Andrea Gregor, and Marco Marelli. 2025. Cracking arbitrariness: A data-driven study of auditory iconicity in spoken English. *Psychonomic Bulletin & Review* 32.1425–42. https://doi.org/10.3758/s13423-024-02630-0.

Drager, Katie K. 2011. Sociophonetic variation and the lemma. *Journal of Phonetics* 39.694–707. https://doi.org/10.1016/j.wocn.2011.08.005.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

Duanmu, San. 2007. *The phonology of Standard Chinese*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780199215782.001.0001.

Elman, Jeffrey L. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science* 33.547–82. https://doi.org/10.1111/j.1551-6709.2009.01023.x.

Ernestus, Mirjam. 2000. *Voice assimilation and segment reduction in casual Dutch: A corpus-based study of the phonology-phonetics interface*. Utrecht: LOT. https://hdl.handle.net/2066/75101.

Firth, John Rupert. 1968. *Selected papers of J. R. Firth, 1952–59*. Bloomington: Indiana University Press.

Fon, Janice. 2004. *A preliminary construction of Taiwan Southern Min spontaneous speech corpus*. Technical Report NSC-92-2411-H-003-050-. Taipei: National Science Council.

Fon, Janice, and Wen-Yu Chiang. 1999. What does Chao have to say about tones?—A case study of Taiwan Mandarin. *Journal of Chinese Linguistics* 27.13–37. https://www.jstor.org/stable/23756742.

Fon, Janice, and Hui-Ju Hsu. 2007. Positional and phonotactic effects on the realization of dipping tones in Taiwan Mandarin. *Phonology and phonetics, tones and tunes, vol. 2: Experimental studies in word and sentence prosody*, ed. by Carlos Gussenhoven and Tomas Riad, 239–69. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110207576.2.239.

Fu, Jo-Wei. 1999. Chinese tonal variation and social network—A case study in Tantzu Junior High School. Taichung: Providence University master's thesis.

Fujisaki, Hiroya. 2004. Information, prosody, and modeling—with emphasis on tonal features of speech. *Proceedings of Speech Prosody 2004*. https://www.isca-archive.org/speechprosody_2004/fujisaki04_speechprosody.pdf.

Gahl, Susanne. 2008. *Time* and *thyme* are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84.474–96. https://doi.org/10.1353/lan.0.0035.

Gahl, Susanne, and R. Harald Baayen. 2024. *Time* and *thyme* again: Connecting English spoken word duration to models of the mental lexicon. *Language* 100.623–70. https://doi.org/10.1353/lan.2024.a947037.

Gahl, Susanne; Yao Yao; and Keith Johnson. 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66.789–806. https://doi.org/10.1016/j.jml.2011.11.006.

Gårding, Eva. 1987. Speech act and tonal pattern in Standard Chinese: Constancy and variation. *Phonetica* 44.13–29. https://doi.org/10.1159/000261776.

Goldman, Jean-Philippe. 2011. Easyalign: An automatic phonetic alignment tool under Praat. *Proceedings of Interspeech 2011*, 3233–36. https://doi.org/10.21437/Interspeech.2011-815.

Günther, Fritz; Luca Rinaldi; and Marco Marelli. 2019. Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science* 14.1006–33. https://doi.org/10.1177/1745691619861372.

Harris, Zellig S. 1954. Distributional structure. *WORD* 10.146–62. https://doi.org/10.1080/00437956.1954.11659520.

Heitmeier, Maria; Yu-Ying Chuang; and R. Harald Baayen. 2021. Modeling morphology with linear discriminative learning: Considerations and design choices. *Frontiers in Psychology* 12:720713. https://doi.org/10.3389/fpsyg.2021.720713.

Heitmeier, Maria; Yu-Ying Chuang; and R. Harald Baayen. 2023. How trial-to-trial learning shapes mappings in the mental lexicon: Modelling lexical decision with linear discriminative learning. *Cognitive Psychology* 146:101598. https://doi.org/10.1016/j.cogpsych.2023.101598.

Heitmeier, Maria; Yu-Ying Chuang; and R. Harald Baayen. 2026. *The discriminative lexicon: Theory, implementation in the Julia package JudiLing, and applications*. Cambridge: Cambridge University Press.

Ho, Aichen T. 1976. The acoustic variation of Mandarin tones. *Phonetica* 33.353–67. https://doi.org/10.1159/000259792.

Howie, John M. 1974. On the domain of tone in Mandarin. *Phonetica* 30.129–48. https://doi.org/10.1159/000259484.

Hsieh, Hsin-I. 1976. On the unreality of some phonological rules. *Lingua* 38.1–19. https://doi.org/10.1016/0024-3841(76)90038-3.

Hsieh, Po-jen. 2013. Prosodic markings of semantic predictability in Taiwan Mandarin. *Proceedings of Interspeech 2013*, 553–57. https://doi.org/10.21437/Interspeech.2013-154.

Hsieh, Shu-Kai, and Yu-Hsiang Tseng. 2020. Tutorial on sense-aware computing in Chinese (version 0.1.6). Paper presented at the 32nd conference on Computational Linguistics and Speech Processing (ROCLING 2020). https://github.com/lopentu/CwnSenseTagger.

Huang, Chu-Ren; Shu-Kai Hsieh; Jia-Fei Hong; Yun-Zhu Chen; I-Li Su; Yong-Xiang Chen; and Shen-Wei Huang. 2010. Constructing Chinese WordNet: Design principles and implementation [in Chinese]. *Zhong-Guo-Yu-Wen* 24(2).169–86.

Huang, Eric; Richard Socher; Christopher Manning; and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long papers)*, 873–82. https://aclanthology.org/P12-1092/.

Huang, Junjie; Duyu Tang; Wanjun Zhong; Shuai Lu; Linjun Shou; Ming Gong; Daxin Jiang; and Nan Duan. 2021. WhiteningBERT: An easy unsupervised sentence embedding approach. *Findings of the Association for Computational Linguistics: EMNLP* 2021, 238–44. https://doi.org/10.18653/v1/2021.findings-emnlp.23.

Huang, Po-Hsuan, and Chenhao Chiu. 2023. Production and perception of coarticulated tones: The cases of Taiwan Mandarin and Taiwan Southern Min. Taipei: National Taiwan University, MS. https://doi.org/10.2139/ssrn.4637487.

Huang, Yi-Hsuan. 2008. Dialectal variations on the realization of high tonal targets in Taiwan Mandarin. Taipei: National Taiwan University master's thesis. http://ntur.lib.ntu.edu.tw//handle/246246/179781.

Iacobacci, Ignacio; Mohammad Taher Pilehvar; and Roberto Navigli. 2015. SensEmbed: Learning sense embeddings for word and relational similarity. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th*

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

International Joint Conference on Natural Language Processing (Volume 1: Long papers), 95–105. https://doi.org/10.3115/v1/P15-1010.

Jin, Xiaoyun; Mirjam Ernestus; and R. Harald Baayen. 2025. A corpus-based investigation of pitch contours of monosyllabic words in conversational Taiwan Mandarin. arXiv:2409.07891 [cs.CL]. https://arxiv.org/abs/2409.07891.

Johnson, Keith. 2004. Massive reduction in conversational American English. *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium*, Tokyo, The National International Institute for Japanese Language, 29–54.

Kendall, David G. 1977. The diffusion of shape. *Advances in Applied Probability* 9.428–30. https://doi.org/10.2307/1426091.

Kilgarriff, Adam. 2007. Word senses. *Word sense disambiguation: Algorithms and applications*, ed. by Eneko Agirre and Philip Edmonds, 29–46. Dordrecht: Springer. https://doi.org/10.1007/978-1-4020-4809-8_2.

Kochanski, Greg, and Chilin Shih. 2003. Prosody modeling with soft templates. *Speech Communication* 39.311–52. https://doi.org/10.1016/S0167-6393(02)00047-X.

Kuhn, Max, and Kjell Johnson. 2013. *Applied predictive modeling*. Dordrecht: Springer. https://doi.org/10.1007/978-1-4614-6849-3.

Ladd, Robert, and Kim E. A. Silverman. 1984. Vowel intrinsic pitch in connected speech. *Phonetica* 41.31–40. https://doi.org/10.1159/000261708.

Landauer, Thomas K., and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104.211–40. https://doi.org/10.1037/0033-295X.104.2.211.

Lee, Ok Joo. 2005. *The prosody of questions in Beijing Mandarin*. Columbus: The Ohio State University dissertation. http://rave.ohiolink.edu/etdc/view?acc_num=osu1122332580.

Levelt, Willem J. M.; Ardi Roelofs; and Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22.1–38. https://doi.org/10.1017/S0140525X99001776.

Li, Qian, and Yiya Chen. 2016. An acoustic study of contextual tonal variation in Tianjin Mandarin. *Journal of Phonetics* 54.123–50. https://doi.org/10.1016/j.wocn.2015.10.002.

Liu, Fang, and Yi Xu. 2005. Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica* 62.70–87. https://doi.org/10.1159/000090090.

Lohmann, Arne. 2018. Cut (N) and cut (V) are not homophones: Lemma frequency affects the duration of noun–verb conversion pairs. *Journal of Linguistics* 54.753–77. https://doi.org/10.1017/S0022226717000378.

Lu, Yuxin; Yu-Ying Chuang; and R. Harald Baayen. 2025a. Form and meaning co-determine the realization of tone in Taiwan Mandarin spontaneous speech: The case of tone 3 sandhi. arXiv:2408.15747 [cs.CL]. https://arxiv.org/abs/2408.15747.

Lu, Yuxin; Yu-Ying Chuang; and R. Harald Baayen. 2025b. The realization of tones in spontaneous spoken Taiwan Mandarin: A corpus-based survey and theory-driven computational modeling. arXiv:2408.15747 [cs.CL]. https://doi.org/10.48550/arXiv.2408.15747.

Ma, Wei-Yun, and Keh-Jiann Chen. 2003. Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff. *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, 168–71. https://doi.org/10.3115/1119250.1119276.

Marsolek, Chad J. 2008. What antipriming reveals about priming. *Trends in Cognitive Science* 12.176–81. https://doi.org/10.1016/j.tics.2008.02.005.

Martinet, André. 1965. *La linguistique synchronique: Études et recherches*. Paris: Presses Universitaires de France.

Moore, Corinne B., and Allard Jongman. 1997. Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America* 102.1864–77. https://doi.org/10.1121/1.420092.

Neelakantan, Arvind; Jeevan Shankar; Alexandre Passos; and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1059–69. https://doi.org/10.3115/v1/D14-1113.

Nieder, Jessica; Yu-Ying Chuang; Ruben van de Vijver; and R. Harald Baayen. 2023. A discriminative lexicon approach to word comprehension, production, and processing: Maltese plurals. *Language* 99.242–74. https://doi.org/10.1353/lan.2023.a900087.

Ouyang, Iris Chuoying, and Elsi Kaiser. 2015. Prosody and information structure in a tone language: An investigation of Mandarin Chinese. *Language, Cognition and Neuroscience* 30.57–72. https://doi.org/10.1080/01690965.2013.805795.

Perek, Florent, and Martin Hilpert. 2017. A distributional semantic approach to the periodization of change in the productivity of constructions. *International Journal of Corpus Linguistics* 22.490–520. https://doi.org/10.1075/ijcl.16128.per.

Pilehvar, Mohammad Taher, and Jose Camacho-Collados. 2020. *Embeddings in natural language processing: Theory and advances in vector representations of meaning*. San Rafael, CA: Morgan & Claypool. https://doi.org/10.1007/978-3-031-02177-0.

Plag, Ingo; Julia Homann; and Gero Kunter. 2017. Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics* 53.181–216. https://doi.org/10.1017/S0022226715000183.

Prom-On, Santitham; Yi Xu; and Bundit Thipakorn. 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. *The Journal of the Acoustical Society of America* 125.405–24. https://doi.org/10.1121/1.3037222.

R Core Team. 2022. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Reisinger, Joseph, and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. *Human language technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 109–17. https://aclanthology.org/N10-1013/.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

Saito, Motoki; Fabian Tomaschek; Ching-Chu Sun; and R. Harald Baayen. 2023. Articulatory effects of frequency modulated by inflectional meanings. *Interfaces of phonetics*, ed. by Marcel Schlechtweg, 125–54. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110783452-005.

Salton, Gerard; Andrew Wong; and Chungshu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18.613–20. https://doi.org/10.1145/361219.361220.

Sampson, Geoffrey. 2015. A Chinese phonological enigma. *Journal of Chinese Linguistics* 43.679–91. https://doi.org/10.1353/jcl.2015.0014.

Sampson, Geoffrey. 2019. An unaddressed phonological contradiction. *International Journal of Chinese Linguistics* 6.221–37. https://doi.org/10.1075/ijchl.19007.sam.

Schütze, Hinrich. 1992. Word space. *Advances in Neural Information Processing Systems 5 (NIPS Conference)*, 895–902.

Shafaei-Bajestan, Elnaz; Masoumeh Moradipour-Tari; Peter Uhrig; and R. Harald Baayen. 2024. The pluralization palette: Unveiling semantic clusters in English nominal pluralization through distributional semantics. *Morphology* 34.369–413. https://doi.org/10.1007/s11525-024-09428-9.

Shen, Xiaonan Susan. 1989. Interplay of the four citation tones and intonation in Mandarin Chinese. *Journal of Chinese Linguistics* 17.61–74. https://www.jstor.org/stable/23757125.

Shen, Xiaonan Susan. 1990a. *The prosody of Mandarin Chinese*. Berkeley: University of California Press.

Shen, Xiaonan Susan. 1990b. Tonal coarticulation in Mandarin. *Journal of Phonetics* 18.281–95. https://doi.org/10.1016/S0095-4470(19)30394-8.

Shen, Xiaonan Susan, and Maocan Lin. 1991. A perceptual study of Mandarin tones 2 and 3. *Language and Speech* 34.145–56. https://doi.org/10.1177/002383099103400202.

Shi, Bao, and Jialu Zhang. 1987. Vowel intrinsic pitch in Standard Chinese. *Proceedings of the 11th International Congress of Phonetic Sciences (ICPhS)*, Tallinn, 142–45. https://www.coli.uni-saarland.de/groups/FK/speech_science/icphs/ICPhS1987/11_ICPhS_1987_Vol_1/p11.1_142.pdf.

Shih, Chilin. 1988. Tone and intonation in Mandarin. *Working Papers, Cornell Phonetics Laboratory* 3.83–109. https://doi.org/10.5281/zenodo.3735401.

Shih, Chilin. 1997. Declination in Mandarin. *Proceedings of Intonation: Theory, Models and Applications*, 293–96. https://www.isca-archive.org/int_1997/shih97_int.html.

Shih, Chilin, and Greg P. Kochanski. 2000. Chinese tone modeling with Stem-ML. *Sixth International Conference on Spoken Language Processing (ICSLP 2000)* 2.67–70. https://doi.org/10.21437/ICSLP.2000-210.

Sun, Ching Chu; Peter Hendrix; Jianqiang Ma; and R. Harald Baayen. 2018. Chinese lexical database (CLD). *Behavior Research Methods* 50.2606–29. https://doi.org/10.3758/s13428-018-1038-3.

Tang, Kevin, and Jason A. Shaw. 2021. Prosody leaks into the memories of words. *Cognition* 210:104601. https://doi.org/10.1016/j.cognition.2021.104601.

Tang, Ping, and Shanpeng Li. 2020. The acoustic realization of Mandarin tones in fast speech. *Proceedings of Interspeech 2020*, 1938–41. https://doi.org/10.21437/Interspeech.2020-1274.

Tseng, Chiu-yu. 1981. *An acoustic phonetic study on tones in Mandarin Chinese*. Providence, RI: Brown University dissertation.

Tseng, Shu-Chuan. 2005a. Contracted syllables in Mandarin: Evidence from spontaneous conversations. *Language and Linguistics* 6.153–80. https://www.ling.sinica.edu.tw/item/en?act=journal&code=download&article_id=153.

Tseng, Shu-Chuan. 2005b. Syllable contractions in a Mandarin conversational dialogue corpus. *International Journal of Corpus Linguistics* 10.63–83. https://doi.org/10.1075/ijcl.10.1.04tse.

van der Maaten, Laurens, and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9.2579–2605. https://jmlr.org/papers/v9/vandermaaten08a.html.

Vulić, Ivan; Edoardo Maria Ponti; Robert Litschko; Goran Glavaš; and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7222–40. https://doi.org/10.18653/v1/2020.emnlp-main.586.

Wang, H. Samuel. 1972. An experimental study on the productivity of Taiwanese tone sandhi. *Proceedings of The Third International Symposium on Language and Linguistics: Pan-Asiatic linguistics*, vol. 1, 116–29. http://sealang.net/sala/archives/pdf8/wang1992experimental.pdf.

Whalen, Douglas H., and Andrea G. Levitt. 1995. The universality of intrinsic f0 of vowels. *Journal of Phonetics* 23.349–66. https://doi.org/10.1016/S0095-4470(95)80165-0.

Wood, Simon N. 2017. *Generalized additive models: An introduction with R*. 2nd edn. New York: Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279.

Wu, E-Chin. 2009. The effect of Min proficiency on the realization of tones and foci in Taiwan Mandarin-Min bilinguals. Taipei: National Taiwan University master's thesis.

Wu, Shu-Juan. 2003. A sociolinguistic study of Chinese tonal variation in Puli, Nantou, Taiwan. Taichung: Providence University master's thesis.

Wu, Yaru; Martine Adda-Decker; and Lori Lamel. 2020. Mandarin lexical tones: A corpus-based study of word length, syllable position and prosodic position on duration. *Proceedings of Interspeech 2020*, 1908–12. https://doi.org/10.21437/Interspeech.2020-1614.

Wu, Yaru; Martine Adda-Decker; and Lori Lamel. 2023. Mandarin lexical tone duration: Impact of speech style, word length, syllable position and prosodic position. *Speech Communication* 146.45–52. https://doi.org/10.1016/j.specom.2022.11.001.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press

Xu, Ching X., and Yi Xu. 2003. Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association* 33.165–81. https://doi.org/10.1017/S0025100303001270.
Xu, Yi. 1994. Production and perception of coarticulated tones. *The Journal of the Acoustical Society of America* 95.2240–53. https://doi.org/10.1121/1.408684.
Xu, Yi. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics* 25.61–83. https://doi.org/10.1006/jpho.1996.0034.
Xu, Yi. 1998. Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica* 55.179–203. https://doi.org/10.1159/000028432.
Xu, Yi. 1999. Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics* 27.55–105. https://doi.org/10.1006/jpho.1999.0086.
Xu, Yi. 2001. Sources of tonal variations in connected speech. *Journal of Chinese Linguistics Monograph Series* 17.1–31. https://www.jstor.org/stable/23825558.
Xu, Yi, and Xuejing Sun. 2002. Maximum speed of pitch change and how it may relate to speech. *The Journal of the Acoustical Society of America* 111.1399–1413. https://doi.org/10.1121/1.1445789.
Zhang, Caicai; Quansheng Xia; and Gang Peng. 2015. Mandarin third tone sandhi requires more effortful phonological encoding in speech production: Evidence from an ERP study. *Journal of Neurolinguistics* 33.149–62. https://doi.org/10.1016/j.jneuroling.2014.07.002.
Zhang, Jie, and Yuwen Lai. 2010. Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology* 27.153–201. https://doi.org/10.1017/S0952675710000060.
Zhang, Ping. 1993. Model selection via multifold cross validation. *The Annals of Statistics* 21.299–313. https://doi.org/10.1214/aos/1176349027.
Zhang, Sheng; P. C. Ching; and Fanrang Kong. 2006. Acoustic analysis of emotional speech in Mandarin Chinese. *International Symposium on Chinese Spoken Language Processing (ISCSLP 2006)*, 57–66. https://www.isca-archive.org/iscslp_2006/zhang06b_iscslp.pdf.
Zhao, Yuan, and Dan Jurafsky. 2009. The effect of lexical frequency and Lombard reflex on tone hyperarticulation. *Journal of Phonetics* 37.231–47. https://doi.org/10.1016/j.wocn.2009.03.002.

Yu-Ying Chuang is an assistant professor in the Department of Taiwan Culture, Languages, and Literature at the National Taiwan Normal University. Email: yuying.chuang@ntnu.edu.tw.

Melanie J. Bell is a professor in the School of Humanities and Social Sciences at Anglia Ruskin University, Cambridge, UK. Email: melanie.bell@aru.ac.uk.

Yu-Hsiang Tseng is a post-doctoral researcher in the Department of Linguistics at the University of Tübingen, Germany. Email: yu-hsiang.tseng@uni-tuebingen.de.

R. Harald Baayen is a professor in the Department of Linguistics at the University of Tübingen, Germany. Email: harald.baayen@uni-tuebingen.de.

Cite this article: Chuang, Yu-Ying; Melanie J. Bell; Yu-Hsiang Tseng; and R. Harald Baayen. 2026. Word-specific tonal realizations in Mandarin. *Language* 102, 1–45. https://doi.org/10.1017/S0097850725000001.

https://doi.org/10.1017/S0097850725000001 Published online by Cambridge University Press