The realization of tones in spontaneous spoken Taiwan Mandarin: a corpus-based survey and theory-driven computational modeling

Yuxin Lu¹, Yu-Ying Chuang², R.Harald Baayen³

 Quantitative Linguistics, Eberhard Karls Universität Tübingen, Tübingen, Germany Email: yuxin.lu@uni-tuebingen.de
 Department of Taiwan Culture, Languages and Literature, National Taiwan Normal University, Taipei, Taiwan Email: yuying.chuang@ntnu.edu.tw
 Quantitative Linguistics, Eberhard Karls Universität Tübingen, Tübingen, Germany Email: harald.baayen@uni-tuebingen.de

April 1, 2025

Abstract

A growing body of literature has demonstrated that semantics can co-determine fine phonetic detail. However, the complex interplay between phonetic realization and semantics remains understudied, particularly in pitch realization. The current study investigates the tonal realization of Mandarin disyllabic words with all 20 possible combinations of two tones, as found in a corpus of Taiwan Mandarin spontaneous speech. We made use of Generalized Additive Mixed Models (GAMs) to model f0 contours as a function of a series of predictors, including gender, tonal context, tone pattern, speech rate, word position, bigram probability, speaker and word. In the GAM analysis, word and sense emerged as crucial predictors of f0 contours, with effect sizes that exceed those of tone pattern. For each word token in our dataset, we then obtained a contextualized embedding by applying the GPT-2 large language model to the context of that token in the corpus. We show that the pitch contours of word tokens can be predicted to a considerable extent from these contextualized embeddings, which approximate tokenspecific meanings in contexts of use. The results of our corpus study show that meaning in context and phonetic realization are far more entangled than standard linguistic theory predicts.

Keywords: contextualized embeddings; Discriminative Lexicon Model (DLM); Generalized Additive Models (GAMs); Mandarin tones; spontaneous speech; word-specific tonal realization

1 Introduction

Mandarin Chinese is a tone language with four lexical tones: a high level tone (T1), a rising tone (T2), a low falling-rising tone known as a dipping tone (T3), and a falling tone (T4). Mandarin Chinese also has a so-called neutral or floating tone (T0), which is often described as unstressed, weaker in intensity, and shorter in duration (Chao, 1968). The present study reports the results of an investigation of the realization of the Mandarin tones in a corpus of Taiwan Mandarin spontanenous speech. We first present our corpus based findings, and then present a theory-driven explanation of our findings using the Discriminative Lexicon Model (Baayen et al., 2019; Heitmeier et al., 2025).

The corpus that we made use of was compiled by Fon (2004), originally with the aim of clarifying the influence of Southern Min on Mandarin Chinese as spoken in Taiwan. In what follows, we refer to this corpus as the Corpus of Spontaneous Taiwan Mandarin. The focus of our study is on the realization in this corpus of the tones in words consisting of two syllables. The tonal realization of disyllabic words has been studied before in laboratory speech. Xu (1997) examined the pitch contours of the 16 combinations of the 4 standard tones realized on the two-syllables /ma-ma/, embedded in carrier sentences, and produced by male speakers of Beijing Mandarin. Factors that are known to co-determine the realization of tones, such as speaking rate and the tones on adjacent words, were carefully controlled for. This study showed that in laboratory speech, the tones of the single-syllable constituents are often somewhat different. For instance, a rising tone followed by a falling tone (T2-T4) was observed to be realized as a fall, followed by a rise, and concluded with a fall.

To our knowledge, there currently are no studies that address the tonal realizations of all tonal combination for disyllabic words in spontaneous conversation. It is well-known that spontaneous speech can differ markedly from formal speech. Given that in spontaneous speech, words are often realized with various reduced forms (see, e.g., Ernestus, 2000; Johnson, 2004; Chung, 2006, for Dutch, English, and Mandarin Chinese, respectively), it is an open question to what extent the canonical four tones of Mandarin are preserved in spontaneous speech. This is one reason why we carried out a detailed investigation of the realization of tone in disyllabic words as found in the Corpus of Spontaneous Taiwan Mandarin. Importantly, we considered not only the 16 combinations of tones studied by Xu (1997), but also the 4 combinations of a standard tone followed by a neutral tone (T1-T0, T2-T0, T3-T0, and T4-T0).

The second reason we carried out this corpus survey is that previous corpus-based research provides strong evidence for the realization of words' tones, as represented by their f0 (pitch) contours, is only in part determined by the canonical tones of the constituent syllables, and that, surprisingly, words' meanings play a much more important role (Chuang et al., 2023, 2024; Lu et al., 2024; Jin et al., 2024) in shaping how the tones are actually realized. In section 2 we provide further details on these findings, and also point to several other studies indicating that meaning and phonetic form are far more entangled than is generally assumed. Here, we note that if indeed fine semantic detail is reflected in fine phonetic detail, this challenges influential axioms of linguistic theory, such as the arbitrariness of the linguistic sign (De Saussure, 1966) and the dual articulation of language (Martinet, 1965).

A recent theory of the lexicon and lexical processing that rejects these axioms is the Discriminative Lexicon Model (Baayen et al., 2019; Heitmeier et al., 2025). This model represents both words' forms and their meanings as high-dimensional numeric vectors, and posits functions that map form vectors onto meaning vectors for comprehension, and meaning vectors onto form vectors (production). In the present study, we zoom in on only a small part of the production process, and ask whether it is possible to start out with a word's meaning vector (using context-specific embeddings from distributional semantics and Large Language Models) and to predict that word's pitch contour using a general mapping from semantic vectors to pitch contours. In section 4, we show that this is indeed possible with an accuracy that is surprisingly far above chance level. We will also show that the canonical tone pattern of a two-syllable word can be predicted from the centroid of the embeddings of the words sharing that canonical tone pattern. Our results raise many questions, for which, as will become clear in the general discussion, we only have tentative answers.

The remainder of this paper is structured as follows. Section 2 introduces the many factors that co-determine how tones are realized, and also provides an overview of previous research on isomorphies between semantics and phonetic realization. Section 3 introduces the corpus that we investigated, and provides details on data pre-processing and the statistical method that we used to analyze the corpus data. Section 3.4 reports our results: across all 20 tone patterns, words' meanings provide a surprisingly good window on their pitch contours. Section 4 presents our theory-driven computational modeling study, showing that token-specific pitch contours can be predicted from token-specific embeddings calculated based on their discourse context. Finally, Section 5 presents our thoughts on the implications of our findings.

2 Semantics and phonetic realization

2.1 Spoken duration and articulation

Evidence is accumulating that subtle differences in meaning can be reflected in the fine phonetic details of how words are actually realized in corpora of natural speech, including aspects such as spoken word (Gahl and Baayen, 2024) duration, segment duration (Plag et al., 2017), and tongue position (Saito, 2024).

Heterographic homophones are words with the same pronunciation but different spellings and meanings, such as *time* and *thyme*. For a long time, homophones were thought to sound identical (see, e.g., Jescheniak and Levelt, 1994). However, Gahl (2008), using the Switchboard corpus (Godfrey et al., 1992) reported that heterographic homophones such as *time* and *thyme* have different acoustic durations, with more frequent homophones (*time*) being pronounced with shorter durations than their less frequent homophonic counterparts (*thyme*). Lohmann (2018) similarly observed that the duration of words such as *cut* depends on whether they are used as nouns or verbs. Both studies explain these effects in terms of how frequency of use affects lexical access in speech production. However, Gahl and Baayen (2024) reported that computational modeling with the Discriminative Lexicon Model provided strong evidence that the meanings of English homophones (represented by embeddings) are strong co-determinants of their spoken word duration, even when word frequency is controlled for. They argued that a powerful predictor of a homophone's spoken word duration is the degree of support it receives from the semantics, such that greater semantic support predicts longer spoken word duration.

In addition to durational differences at the word level, durational differences have also been observed at the phonemic level in corpus studies, particularly for the realization of word-final /s/ or /z/ (henceforth referred to as S) in English. In the Buckeye corpus (Pitt et al., 2005), word-final S has been found to vary in duration depending on its morphological function: non-morphemic S is pronounced longer than plural S, which, in turn, is pronounced longer than clitic S (Plag et al., 2017; Tomaschek et al., 2021; Zimmermann et al., 2016). Furthermore, Plag et al. (2020) found that genitive plural S showed significantly longer durations than plural S. Schmitz (2022) utilized a pseudoword paradigm to demonstrate that the morphological category of word-final S (non-morphemic > plural > clitics) influences its phonetic realization.

The relationship between semantics and phonetic realization has also been demonstrated beyond durational differences. Drager (2011) found that the pronunciation of the English 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. Furthermore, in line with Gahl and Baayen (2024), Saito (2024) and Saito et al. (2024) reported for the KEC corpus of German spontaneous speech (Arnold and Tomaschek, 2016), which also registers tongue movements using electromagnetic articulography, greater semantic support leads to a lower position of the tongue tip for the vowel /a/, indicating hyperarticulation.

2.2 Tone in Mandarin Chinese

The preceding section reviewed recent evidence that semantics and phonetic realization are entangled to a greater extent than has often been assumed. In this section, we zoom in on how word meaning affects the realization of tone. To do so, we first need to provide some further background on the factors that have already been reported to co-determine the realization of tone.

It is well known that the way in which tones are realized in connected speech differs from their canonical realization. How tones are realized has been described as depending on the properties of the segments in the syllable that carries a given tone (Ho, 1976b; Ohala and Eukel, 1976; Xu and Xu, 2003). Tonal variation in connected speech has also been reported to be shaped by the tones of adjacent words (tonal coarticulation Xu, 1997), by speaking rate (Xu and Sun, 2002), by sentence intonation (Ho, 1976b; Wu et al., 2020) and by a speaker's individual speaking style (Stanford, 2016).

At the socio-geographic level, cross-dialectal research has reported that different varieties of Mandarin exhibit varying tone inventories (Chang, 2010; Zhao, 2023). For the present study, we note that in Standard Mandarin, the realization of a neutral tone following a given lexical tone has been reported to be largely determined by this preceding tone. Furthermore, the f0 contour of a neutral tone has been claimed to approach a low pitch target by the end of the carrying syllable (Xu, 2024). However, in Taiwan Mandarin, the behavior of the the neutral tone has been reported to be different. It can either be indistinguishable from one of the four canonical lexical tones, or it can be realized as a static mid-low pitch target (Huang, 2018).

From the above overview, it will be clear that the realization of tone is co-determined by a multitude of different factors. A newcomer in this arena is word meaning. Chuang et al. (2024) studied the pitch contours of di-syllabic words with an initial rising tone followed by a falling tone (henceforth the T2-T4 tone pattern). This study used a generalized additive model (GAM Wood, 2017) to decompose an observed pitch contour into separate pitch contours, capturing the effects over time of predictors such as speech rate, neighboring tones, and segmental properties. What Chuang et al. (2024) report is that the GAMs provide strong support for word-specific pitch contour components, while controlling for other variables such as segments, gender, speaker, speech rate, and the tones of adjacent words. They also show that the statistical evidence is even stronger for sense-specific pitch contours, which suggests that these effects are semantic in nature. The importance of words' meanings has been replicated for Mandarin disyllabic words with T2-T3 and T3-T3 tone pattern by Lu et al. (2024), and for monosyllabic Mandarin words by Jin et al. (2024). For di-syllabic words with T2-T3 and T3-T3 tone pattern, the variable importance of words' meanings was on a par of that of tonal context (tone sandhi), the other most important predictor of words' pitch contours.

To illustrate the challenges that an analysis of tones in natural speech has to face, consider Figure 1, which displays the f0 contours of a selection of tokens of word types with a falling tone followed by a rising tone (the T4-T2 tone pattern) in the Corpus of Spontaneous Taiwan Mandarin (Fon, 2004). In the left panel, the pitch contours of six tokens from different word types are presented. All words have the same canonical tone pattern: T4-T2. Token XMC_GY_4119_不能 (*bu4neng2*, 'cannot') (indicated by light blue) shows an initial sharp f0 rise, followed by a fall, and then a shallow rise. Token XMC_GY_8107_問題 (*wen4ti2*, 'problem') (indicated by purple) has a much lower initial f0 than the other tokens. The initial f0 of token XMC_GY_1025_後來 (*hou4lai2*, 'later') (indicated by red) is unavailable due to the unvoiced initial /h/. In the right panel of Figure 1, we present four tokens of the same word type 幸福 (*xing4fu2*, 'happiness'). One of its tokens is also presented in the left panel (indicated by yellow). The four tokens of 幸福 also exhibit considerable variability in their f0 contours.

In what follows, we take on the challenge of modeling the realization of tone, taking into account the many factors reported to co-determine pitch contours, such as gender, speaker, neighboring tones, speech rate, word position, and bigram probability. Following up on earlier work (Chuang et al., 2024; Jin et al., 2024; Lu et al., 2024), the 'pitch signatures' of individual words are of primary interest. Two hypotheses guide our research.



Figure 1: A selection of tokens in spoken Taiwan Mandarin. Left panel: six tokens representing six different word types, all sharing the tone pattern T4-T2 (a falling tone followed by a rising tone). The tokens are 後來 (*hou4lai2*, 'later'), 幸福 (*xing4fu2*, 'happiness'), 去年 (*qu4nian2*, 'last year'), 不能 (*bu4neng2*, 'cannot'), 自然 (*zi4ran2*, 'nature'), 問題 (*wen4ti2*, 'problem'). Right panel: four tokens representing the word type 幸福 (*xing4fu2*, 'happiness'). All f0 contours shown here are produced by the same speaker.

- 1. The meanings of words co-determine the phonetic details of how the tones of these words are produced.
- The pitch contours of word tokens as found in spontaneous Mandarin conversations can be predicted from token-specific meaning vectors with above-chance accuracy using computational modeling.

In the next section, we describe the data that we collected from the Corpus of Spontaneous Taiwan Mandarin, and present our statistical analyses. Section 4 complements this exploratory part of our study with theory-driven computational modeling.

3 Data collection and statistical analysis

3.1 The corpus

The data used in the present study come from the Taiwan Mandarin Spontaneous Conversation Corpus (Fon, 2004). This corpus contains 30 hours of speech from 55 native speakers of Taiwan Mandarin, with 31 females and 24 males (aged between 20 and 60 years old). In unstructured interviews, participants were encouraged to speak freely, instead of being guided by a standardized set of questions. As a result, this corpus consists of naturally occurring speech with a diverse and varied set of words across speakers.

The corpus was transcribed in traditional Chinese characters at the word level. The speech data were segmented at both the syllable and word levels. Forced alignment was first performed, and the results were later manually reviewed by native Taiwan Mandarin speakers with a background in phonetics. In the current study, we followed the transcriptions and segmentation provided in the corpus.

3.2 Data selection

Disyllabic words with the 20 tone patterns were extracted for analysis (see column 1 in Table A.1). The original dataset comprises 93,701 tokens, representing 7,526 unique word types. Table A.1 presents the counts of tokens and word types associated with each tone pattern. Among these, the T4-T4 pattern is the most frequent in disyllabic words, both token-wise and type-wise. The four tone patterns featuring a neutral tone in the second syllable (T1-T0, T2-T0, T3-T0, and T4-T0) are represented by the lowest numbers of types. There are also relatively few words with T3 (see Wu et al., 2021, for similar observations for journalistic speech).

Table 1: Number of tokens and words grouped by tone pattern in the conversational Taiwan Mandarin corpus.

	Tone pattern	Tokens	Word types	Examples
1	T1-T1	3501	459	應該 (ying1gai1, 'should')
2	T1-T2	3725	458	當然 (dang1ran2, 'of course')
3	T1-T3	2313	333	根本 (gen1ben3, 'at all')
4	T1-T4	7524	706	接觸 (jie1chu4, 'to touch')
5	T1-T0	3034	83	他們 (talmen0, 'they')
6	T2-T1	2763	286	其他 (qi2ta1, 'others')
7	T2-T2	3043	369	同學 (tong2xue2, 'classmate')
8	T2-T3	4539	249	結果 (jie2guo3, 'result')
9	T2-T4	9237	687	學校 (xue2xiao4, 'school')
10	T2-T0	7010	64	什麼 (<i>shen2me0</i> , 'what')
11	T3-T1	2655	252	老師 (lao3shi1, 'teacher')
12	T3-T2	3465	289	感覺 (gan3jue2, 'feeling')
13	T3-T3	3896	276	了解 (liao3jie3, 'to know')
14	T3-T4	7256	595	可是 (ke3shi4, 'but')
15	T3-T0	3295	50	我們 (wo3men0, 'we')
16	T4-T1	3007	400	那些 (na4xie1, 'those')
17	T4-T2	3978	451	後來 (hou4lai2, 'later')
18	T4-T3	3302	419	父母 (fu4mu3, 'parents')
19	T4-T4	13174	989	社會 (she4hui4, 'society')
20	T4-T0	2984	111	爸爸 (ba4ba0, 'daddy')
	Total	93701	7526	

Subsequently, we extracted the sound files of these disyllabic words and measured their f0 values using the *To Pitch* (*cc*) command in Praat (Boersma and Weenink, 2020). For female speakers, the pitch floor was set at 75 Hz and the pitch ceiling at 400 Hz. For male speakers, the pitch floor was set at 50 Hz and the pitch ceiling at 300 Hz. The time step was set to 0.001 seconds, and a Gaussian window was used for optimal F0 estimation. The *To PointProcess* command was then applied to identify the time points of glottal pulses in the voiced sections, from which the corresponding F0 values were extracted. No f0 values were returned when there was no vocal fold vibration due to the presence of voiceless plosives or fricatives, or when creaky voice occurred.

Words with fewer than six tokens were excluded from our dataset, to ensure that each word type had a sufficient number of tokens for analysis. For high-frequency words with more than 200 tokens, we randomly sampled 200 tokens, to prevent model predictions from being biased towards high-frequency words. Furthermore, words contributed by only female speakers or only by male speakers were excluded. This ensured that the tokens of a given word type were contributed by at least two speakers, preventing bias from one speaker's specific way of speaking.

Lastly, tokens with f0 extraction errors were excluded from analysis. These errors typically resulted from pitch halving or doubling. We calculated, for each token, the standard deviation of the differences between consecutive measurements. A large standard deviation indicated high likelihood of discontinuous f0 measurements with abrupt fluctuations. Tokens with a standard deviation greater than the 9th decile of the distribution were considered outliers.

3.3 Predictors

The response variable of interest is f0. We log-transformed f0 to obtain a response variable that approximately follows a Gaussian distribution. As our interest is in production rather than comprehension, we did not make use of modifications of the logarithmic transformation such as the MEL or BARK scales, which are optimized for human perception. The predictors for log f0 are as follows.

- normalized_t For each token, time was normalized between 0 and 1, enabling the modeling of tokens with varying durations on a common scale. Since f0 values were measured every 15 ms, tokens with longer durations have more measurements and, consequently, more data points within the [0,1] interval of normalized time.
- gender A categorical variable with two levels—female and male. Due to physiological differences, female speakers generally produce speech at a higher pitch than male speakers. gender is included as a control variable.
- speaker A factor with anonymized speaker identifiers as levels, required for fine-tuning differences in speakers' height of voice.
- tone_pattern The tonal pattern of the token, as listed in the *tone pattern* column in Table A.1.
- tonal_context preceding_tone is the tone of the syllable immediately preceding a token.
 following_tone is the tone of the syllable immediately following a token. If a pause occurs immediately before or after the token, it is coded as PAUSE. Thus, both preceding_tone
 and following_tone include six possible values: 1, 2, 3, 4, 0, and PAUSE. To represent
 the different tonal contexts in which the token may appear, we define tonal_context as
 the interaction of preceding_tone and following_tone, resulting in a factor with 36
 levels.
- speech_rate Local speech rate, defined as the number of syllables per second for a given token, is calculated over a window extending four characters to the left and four characters to the right of the token. This measurement of speech rate is included as a covariate to control for potential effects of durational differences. To avoid concurvity, duration is therefore not included alongside speech_rate as a predictor, as the two variables are moderately correlated r = -0.55.
- norm_utt_pos Normalized position in the utterance represents the relative position of a word within its utterance. It is calculated by dividing the word's position by the total number of syllables in the utterance, resulting in a value normalized on a scale from 0 to 1. Higher values indicate that the token occurs closer to the end of the utterance. For single-word utterances, the position is coded as 1. Previous research has shown that utterance-final words tend to exhibit a rising pitch (Shih, 2000).

bg_prob_prev Bigram probability quantifies how predictable a word is in its context. This measure

of contextual predictability is based on the relative frequency of the word's co-occurrence with surrounding words. A higher bigram probability indicates that the target word is more predictable within its given context. In general, higher predictability is associated with shorter word durations and greater spectral reduction (Arnon and Priva, 2014; Tang and Bennett, 2018). There is also some evidence showing that contextual predictability influences f0 production, as observed in English (Turnbull, 2017), Taiwan Mandarin (Hsieh, 2013), and Taiwan Southern Min (Wang, 2024). In the present study, following Gahl (2008), bg_prob_prev is calculated as the probability of the occurrence of the target word given the preceding word.

- bg_prob_fol This measure represents the bigram probability of the following word, calculated as the probability of the occurrence of the target word given the following word.
- word A factor with orthographic words, as available in the corpus, as levels. For instance, the token XMC_GY_8107_問題 is coded as 問題 using traditional Chinese characters. The dataset contains 313 unique words, so there are 313 corresponding levels for word.
- sense_type A word can have multiple senses, which are identified based on the contexts in which the word occurs. We used a word sense identification system, described in Hsieh et al. (2024), that utilizes BERT in combination with the Chinese WordNet (Huang et al., 2010).

Of the above list of predictors, the factor tonal_context poses a special challenge for the analysis. tonal_context provides information about the preceding and following tones. Due to the pervasiveness of tonal co-articulation, it is highly probable that the effect of tonal_context varies with the tone pattern of the target word. For example, a preceding high tone will have an effect on a word-initial dipping tone that differs from its effect on a word-initial rising tone. Accounting for such co-articulation is essential for modeling f0 in connected speech. In principle, one could introduce a variable that represents the interaction between tonal_context and tone_pattern (cf. Jin et al., 2024). However, for our dataset, this would result in a variable with 720 levels that is strongly confounded with word and sense_type.

Therefore, we opted to fit separate regression models for f0 across four different most frequent tonal contexts in our dataset, excluding any contexts that involved a "pause" in the preceding or following tone, i.e., the contexts 4.4, 3.4, 4.1, and 4.0 (cf. Table 2). We chose not to include tonal contexts involving a "pause", for two reasons. First, when a tone is preceded or followed by a pause, several context-related variables, such as norm_utt_pos, bg_prob_prev, and bg_prob_fol, are missing, leading to data loss. Second, pauses in speech often signal utterance boundaries, hesitations, or breaths, making the "pause" category inherently heterogeneous.

As shown in Table 2, the final dataset contains 4,283 tokens representing 313 unique word types. The minimum number of tokens per word type is 5, and the maximum is 56. On average, each word type was produced by 9.37 different speakers (range: 2 to 30). Additionally, each speaker contributed an average of 53.31 different word types (range: 4 to 119). For each tonal context, all 20 tone patterns are represented.

3.4 Statistical analysis

3.4.1 Models with word as predictor

The Generalized Additive Model (GAM, Wood, 2017) was used for the statistical analyses, with the bam() function from mgcv package (Wood, 2017) implemented in R (Team, 2020). Four GAMs were fitted to each of the four datasets, using the same model specification:

Tonal context	Number of tokens	Number of word types	Number of tone patterns
4.4	1,794	288	20
3.4	888	240	20
4.1	874	250	20
4.0	727	210	20
Total	4,283	313	20

Table 2: Overview of the four sub-datasets grouped by the four tonal contexts.

logf0 ~ gender +
 s(normalized_t, by=gender,k=4) +
 s(speaker, bs='re') +
 s(normalized_t, tone_pattern, bs='fs', m=1)+
 s(normalized_t, word, bs='fs', m=1) +
 s(speech_rate, by=gender, k=4) +
 ti(normalized_t, speech_rate, k=c(4,4)) +
 s(norm_utt_pos, k=4) +
 ti(normalized_t, norm_utt_pos, k=c(4,4)) +
 s(bg_prob_prev, k=4)+
 ti(normalized_t, bg_prob_prev, k=c(4,4))+
 s(bg_prob_fol, k=4)+
 ti(normalized_t, bg_prob_fol, k=c(4,4))

To account for differences in average pitch height between genders, we included gender as a fixed effect. We added a by-gender thin plate regression smooth of normalized_t, which allows us to capture differing relationships between normalized time and f0 across genders. Other continuous variables, including speech_rate, norm_utt_pos, bg_prob_prev, and bg_prob_fol were likewise modeled with thin plate regression splines. Interactions of covariates with normalized time were modeled with tensor product smooths, using the ti() function.

Furthermore, random intercepts were requested for speaker to account for individual variability in pitch height by speaker. Other discrete variables, including tone_pattern and word, were modeled using factor smooths (nonlinear random effects).

We implemented an AR(1) process (first-order auto-regressive model) in the residuals to take into account the auto-correlations in the time series of pitch measurements. The inclusion of the AR(1) process with an auto-correlation coefficient of rho = 0.95 effectively removed nearly all autocorrelation from the residuals. Summaries of the four models are provided in the Appendix.

Akaike's Information Criterion (AIC) was used to assess variable importance. Figure 2 shows the increase in AIC (indicating a lower-quality fit to the data) resulting from withholding individual predictors from the model specification. A greater increase in AIC when a predictor is excluded suggests a higher importance of that predictor in the model. As shown in Figure 2, across all tonal contexts, withholding the predictor word leads to a substantial increase in AIC scores, ranging from 7430.30 to 12320.26. The increase in AIC when word is omitted from the model specification substantially exceeds the corresponding change observed for any other predictor.

Surprisingly, withholding tone_pattern has a small impact on the model fit with increases in AIC of around 22.66 units (22.66 units for 4.4, 7.33 units for 3.4, 6.34 units for 4.1, and 16.9 units for 4.0). One possible explanation is that word is nested within tone_pattern. When word is removed from the best-fit GAM, withholding tone_pattern results in a larger AIC increase by 7354.08 units for 4.4, 4069.89 units for 3.4, 3026.09 units for 4.1 and 3451.28 units for 4.0. This suggests that tone_pattern still contributes to the model fit, though not as strongly as word. When word is included in the model, the effect of tone_pattern is overshadowed by the stronger effect of word.



Figure 2: The increase in AIC scores when a predictor is withheld from the best-fit model. The AIC increase when word or tone_pattern is withheld is shown in red, and the increase for other predictors is shown in blue. Panels 1 to 4 represent four GAMs with tonal contexts 4.4, 3.4, 4.1, and 4.0, respectively.

Concurvity, analogous to collinearity in linear regression, measures how much a predictor's effect can be explained by other predictors in the model. Concurvity scores range from 0 to 1, with lower values indicating that the contribution of a predictor is less confounded with the contributions of other predictors. As shown in Figure 3, concurvity scores follow a similar pattern for all four GAMs, with the lowest concurvity scores for word. speaker also has relatively low concurvity.



Figure 3: Concurvity scores for selected terms in the four GAMs. The concurvity scores for word and tone_pattern are shown in red, and those for other predictors are shown in blue. From left to right, it presents tonal context 3.4, 4.0, 4.1, and 4.4 respectively. Concurvity scores were calculated based on the best-fit GAMs with all predictors included.

By contrast, the predictor tone_pattern exhibits extremely high concurvity, ranging from 0.998 to 1. This is due to tone pattern being fully predictable given the word. When word is excluded from the model, the concurvity of tone_pattern drops substantially to 0.09. This indicates that word captures information about the word's tone pattern, so when word is included in the model, the tone pattern is also included implicitly. However, if only tone_pattern is specified, word-specific information is not available. This results in a substantially worse fit, which aligns with the AIC change discussed in preceding subsection.

Finally, we note that the by-gender smooths for time (normalized_t:female and normalized _t:male) show very high concurvity — unsurprisingly, as the tonal contours for both genders are highly similar (see Figure 4). Without accounting for other effects, these general contours primarily reflect the pure influence of time on pitch, illustrating how pitch contours change over time. The overall curves exhibit falling contours, which suggests a general declination trend in pitch contours for disyllabic words.



Figure 4: The partial effect of general smooth for the normalized_time for female and male speakers, in different tone contexts. The orange curves indicate the general contours for female speakers, and the blue curves indicate the general contours for male speakers. Vertical grey dashed lines indicate the average syllable boundary, and the horizontal grey dashed line represents the y=0 reference line.

Figure 5 illustrates the partial effects of the 20 tonal patterns across the four tonal contexts under investigation, using color coding to distinguish between the tonal contexts. Within each panel, the

various blue curves represent specific tone patterns associated with different tonal contexts. For example, the lightest blue curve in the upper-left panel represents the T4-T0 tone pattern in the 3.4 tonal context. In other words, it represents a tonal sequence T3-T4-T0-T4, as in the phrase $\pi i \equiv \underline{\mathbb{E}}$ (*you3zhe4me0zhong4*, '...is this heavy'). The red curves were obtained from averaging the four blue curves representing tone patterns under different tonal contexts. The deviations of the blue curves from the corresponding red curves highlight how the actual realizations of a tonal pattern in context differ from the expected effect of tone pattern, irrespective of context.



Figure 5: The effect of tone pattern. The blue curves represent the partial effects of the factor smooth for tone_pattern, combined with the general smooth of normalized_t for female speakers, based on the best-fit models that include the word effect. There is one GAM for each tonal context, resulting in four blue curves representing, in a given panel, the four tonal contexts. The red curves present the mean f0 contours of a tone pattern, calculated by averaging the four f0 contours across the tonal contexts. Thus, the blue curves in each panel illustrate how the tonal context modulates the general curve shown in red. Vertical grey dashed lines indicate the average syllable boundary, and the horizontal grey dashed line represents the y=0 reference line.

For most of the tone patterns, the effects of the neighboring tones on the pitch contour are relatively modest, with as glaring exceptions the T2-T0 tone patterns in the 4.0 tonal context. This tonal sequence T4-T2-T0-T0 (e.g., 對孩子的, *dui4hai2zi0de0*, "for children's ...") shows an unexpectedly low f0. This is probably due to the fact that this tonal sequence is underrepresented in the dataset, with only 9 tokens representing 4 unique word types (cf. Table A.1 in Appendix 1.). However, the effects of tonal context are less pronounced in tone patterns featuring the neutral tone. For tone patterns T1-T0, T2-T0, T3-T0, and T4-T0, the general trend appears to approach a similar mid-low pitch target at the end of the syllable, regardless of the following tone.

For the 3.4 tonal context, 14 of the tonal patterns begin with the lowest f0. This may be a straightforward consequence of Tone 3 being often realized as a low tone in Taiwan Mandarin (Fon and Chiang, 1999). For the 4.0 tonal context, by the end of the word, the f0 tends to be the lowest across all panels. This is probably due to the general curve of female speakers in 4.0 tonal context. The female curve of 4.0 tonal context has a particularly salient falling trend (cf. Figure 4). Apparently, the following neutral tone is magnifying the final downward inclination observed in the vast majority of tone patterns.

Figure 6 displays a selection of predicted pitch contours estimated by the factor smooth for word, combined with the partial effects of the factor smooth for tone_pattern. All words presented here follow the T4-T2 tone pattern in the 4.4 tonal context (i.e., a tonal sequence T4-T4-T2-T4). For instance, this sequence occurs in the phrase 就變成興趣 (*jiu4bian4cheng2xing4qu4*, 'then become an interest'). These partial effects exclude the general intercept and do not account for pitch differences between female and male speakers.

The red dashed curves represent the partial effect of the factor smooth for tone_pattern only, without incorporating the word-specific pitch contours, and are shown to provide a reference for assessing the word-specific effects.

It can be observed that the pitch contours of 後來 (*hou4lai2*, 'later'), 不然 (*bu4ran2*, 'otherwise'), and 不能 (*bu4neng2*, 'cannot') closely align with the predicted tone pattern but are overall shifted upward. Similarly, the pitch contour of 認為 (*ren4wei2*, 'to believe') also follows a similar shape but is shifted downward. However, other words, such as 幹嘛 (*gan4ma2*, 'What for?'), 目前 (*mu4qian2*, 'at present'), and 化學 (*hua4xue2*, 'chemistry'), largely deviate from the general tone pattern. Two words beginning with $\overline{\wedge}$ (*bu4*, expressing negation), $\overline{\wedge}$ 然 (*bu4ran2*, 'otherwise') and $\overline{\wedge}$ 能 (*bu4neng2*, 'cannot'), have very similar contours that run parallel to the contour of the T4-T2 pattern. However, $\overline{\wedge}$ 行 (*bu4xing2*, 'not okay'), displays a steeper fall.

The deviation of the blue curves from the red dashed curves reflects the differences between the predicted pitch contours and the general tone pattern. The word-specific tonal realizations observed here are similar to those reported for Mandarin disyllabic words with T2-T4 tone patterns (Chuang et al., 2024), as well as words with the T2-T3 and T3-T3 tone patterns (Lu et al., 2024).

3.5 Sense-specific tonal realization

In the preceding section, we documented that the tonal realization of Mandarin di-syllabic words varies systematically by word. It is possible that words' segmental make-up is the crucial factor. Alternatively, it is theoretically possible that it is words' meanings that shape their pitch contours, just as in English, the duration of homophones is to a considerable extent co-determined by their meanings (Gahl and Baayen, 2024). If this hypothesis is on the right track, then word sense should be a more precise predictor than word identity. In the following analysis, we explore whether we can replicate previous studies in which sense emerged as an even better predictor of disyllabic words' pitch contours than the word itself (Chuang et al., 2024; Lu et al., 2024). If we can show that a word with different meanings exhibits varying pitch realizations, this will provide further evidence that words' meanings co-determine tonal realization.

In order to explore the value of this hypothesis, we make use of the fact that our data are taken from a corpus, and not a word list. As a consequence, we can estimate a word token's most likely sense in the exact context in which it was used. To determine these most likely senses in context, we made use of the sense identification system proposed by Hsieh et al. (2024), which uses BERT in combination with the Chinese WordNet (Huang et al., 2010). For example, this system identifies the word 先生 (*xian1sheng1*, 'husband, sir') as 'a woman's spouse in a marital relationship' in sentences such as 我先生認為 (*wo3xian1sheng1ren4wei2*, 'My husband thinks ...') or 我先生去睡



Figure 6: A selection of predicted f0 contours for words with the T4-T2 tone pattern. These contours are estimated by combining the partial effects of the factor smooth for word and the corresponding factor smooth for tone_pattern (T4-T2). The dashed red curve represents the partial effect of the T4-T2 tone pattern alone, which is identical across all panels. Vertical grey dashed lines indicate the average syllable boundary, while the horizontal grey dashed line represents the y=0 reference line.

覺 (wo3xian1sheng1qu4shui4jiao4, 'My husband went to sleep ...'). It assigns 先生 (xian1sheng1, 'husband, sir') the sense 'a man addressed in a social context' to when it appears in the phrase 那位 先生 (na4wei4xian1sheng1, 'That gentleman over there ...').

Since not all words in the dataset could be assigned a sense, we excluded words for which no sense type was identified. Second, we removed sense types represented by fewer than six tokens to ensure that each sense type had sufficient data for meaningful analysis. To prevent the model's predictions from being biased toward high-frequency sense types, we limited the maximum number of tokens per sense type. Specifically, for any sense type represented by more than 50 tokens, we randomly sampled 50 tokens from all tokens. We then grouped the dataset by tonal context, as in the previous analysis, resulting in four sub-datasets (see Table 3). The final dataset consists of 3,525 tokens representing 290 unique sense types. After the trimming process, 252 unique word types remain from the initial 313. All 20 tone patterns are present for each tonal context. The distribution of sense types, and word types follow the similar pattern as in the dataset shown in Table A.1.

Tonal context	Tokens	Sense types	Word types	Tone patterns
4.4	1512	266	233	20
3.4	740	220	195	20
4.1	716	228	200	20
4.0	557	171	157	20
Total	3525	290	252	20

Table 3: Overview of trimmed datasets grouped by the four tonal contexts, for the sense analysis.

For the sense analysis, we replaced the factor smooth for word with a factor smooth for sense_type, while keeping all other predictors from the previous analysis.

```
logf0 ~ gender +
    s(normalized_t, by=gender,k=4) +
    s(speaker, bs='re') +
    s(normalized_t, tone_pattern, bs='fs', m=1)+
    s(normalized_t, sense_type, bs='fs', m=1) +
    s(speech_rate, by=gender, k=4) +
    ti(normalized_t, speech_rate, k=c(4,4)) +
    s(norm_utt_pos, k=4) +
    ti(normalized_t, norm_utt_pos, k=c(4,4)) +
    s(bg_prob_prev, k=4)+
    ti(normalized_t, bg_prob_prev, k=c(4,4))+
    s(bg_prob_fol, k=4)+
    ti(normalized_t, bg_prob_fol, k=c(4,4))
```

An AR(1) process in the errors was also included to account for the autocorrelation in the pitch time series. The model summary is available in the Appendix.

To assess the relative importance of sense_type, word, and tone_pattern, we compared three additional models with different predictor structures: (1) a model with sense_type + tone_pattern, (2) a model with word + tone_pattern, and (3) a model sense_type by itself. Table 4 presents the AIC differences resulting from changing or withholding the given variable, relative to the (sense_type + tone_pattern) model.

First consider the GAMs where sense_type was replaced by word. In the case of the 4.4 tonal context, replacing sense_type with word (comparing row 1 and row 2) led to a substantial AIC increase of 457.28 units. This suggests that sense_type is a stronger predictor than word for modeling f0 contours.

Second, comparing row 1 and row 3, removing tone_pattern while retaining sense_type led to a smaller AIC increase of 28.08 units. This indicates that tone_pattern contributes to the model fit, albeit with a relatively minor effect when sense_type is included. A similar AIC pattern across the 3.4, 4.1, and 4.0 tonal contexts further reinforces the stronger influence of sense_type over word in modeling f0 contours.

Tonal Context	Model	AIC	AIC Difference
4.4	all other predictors + sense_type + tone_pattern	-226847.29	_
4.4	all other predictors + word + tone_pattern	-226390.01	457.28
4.4	all other predictors + sense_type	-226824.40	22.89
3.4	all other predictors + sense_type + tone_pattern	-117518.90	-
3.4	all other predictors + word + tone_pattern	-116923.04	595.85
3.4	all other predictors + sense_type	-117515.43	3.47
4.1	all other predictors + sense_type + tone_pattern	-113771.84	-
4.1	all other predictors + word + tone_pattern	-113177.81	594.03
4.1	all other predictors + sense_type	-113765.18	6.66
4.0	all other predictors + sense_type + tone_pattern	-93868.50	-
4.0	all other predictors + word + tone_pattern	-93650.04	218.46
4.0	all other predictors + sense_type	-93861.31	7.20

Table 4: AIC scores for models with different structures of word, sense_type, and tone_pattern, fitted separately to datasets for the four tonal contexts.

Figure 7 displays the predicted tonal contours for different sense types of 另外(*ling4wai4*, 'in addition'), calculated by combining the partial effects of sense_type and tone_pattern. Similar to the red dashed curves in Figure 6, the red dashed curves in Figure 7 again represent the general tone pattern, which is T4-T4 in this case. The three sense types of 另外(*ling4wai4*, 'in addition') are:

'others' (sense1), 'totally different' (sense2), 'in addition to' (sense3). The word 另外 (*ling4wai4*, 'in addition') exhibits clear variations across the three sense types compared to the general tone pattern. The pitch contours of sense 1 (shown in purple) are generally shifted below the general tone pattern, while those of sense 2 (shown in blue) are shifted above it. The pitch contour of sense 3 (shown in yellow) displays two rises, as in the general tone pattern, but is shifted upwards.



Figure 7: A selection of the predicted f0 contours for different sense_type of 与外 (*ling4wai4*, 'in addition') across tonal contexts. The predicted pitch contours represent the partial effect of the factor smooth for sense_type, combined with the corresponding factor smooth for tone_pattern (T4-T4 in this case). The red dashed curves represent the partial effect of T4-T4 tone pattern alone, averaged across four tonal contexts, so the red dashed curve is the same across all panels. Vertical dashed lines indicate the average syllable boundary, and the horizontal grey dashed line represents the y=0 reference line.

3.6 Summary

This section addressed our first hypothesis, namely, that the meanings of words co-determine the phonetic details of how the tones of these words are produced. Our results show that word emerged as a more powerful predictor than all other predictors. Surprisingly, the variable importance of word was substantially greater than that of tone_pattern. The strong effect of word that we observed is line with the results of Jin et al. (2024) and Lu et al. (2024). Jin et al. (2024) also observed, albeit for monosyllabic words, that word was a stronger predictor than tone_pattern. In the study by Lu et al. (2024), however, the variable importance of word was similar to that of tone_pattern.

A further analysis clarified that sense_type is an even better predictor of pitch contours than word. The substantial improvement in model fit contributed by sense_type provides further support for the hypothesis that it is words' meanings that co-determine the fine detail of their pitch contours, replicating the findings of earlier studies (Chuang et al., 2024; Jin et al., 2024; Lu et al., 2024).

4 Theory-driven computational modeling

In this section, we turn to our second hypothesis, exploring whether the tonal realization of a given token can be predicted with reasonable accuracy based on its context-specific meaning using computational modeling. To do so, we make use of the general conceptual framework of the Discriminative Lexicon Model (DLM Heitmeier et al., 2025), a computationally implemented theory that was developed independently of the present data, but that turns out to provide exactly the right approach to predict tonal realization from semantics.

In the introduction, we already explained that the DLM seeks to predict words' forms from their meanings. Both forms and meanings are represented by numeric vectors, and in the simplest possible

set-up, a linear mapping transforms a meaning vector into a form vector (for mappings using deep neural networks, see Heitmeier et al., 2025). For the present study, we are not interested in predicting full word forms, but rather words' pitch contours. What we need, then, are numerical representations of the present Mandarin word tokens' pitch contours on the one hand, and their meanings on the other hand. Following Chuang et al. (2024), we represent words' forms using fixed-length vectors representing pitch contours, and we represent words' meanings using contextualized embeddings obtained with the GPT-2 transformer technology. Importantly, both the pitch vectors and the semantic vectors are context-specific, and thus vary from word token to word token. Chuang et al. (2024) demonstrated that the tonal contours of a given token with T2-T4 tone pattern can be predicted from its context-specific meaning with above-chance accuracy using a linear mapping. In what follows, we consider whether this result generalizes to all 20 tone patterns attested for two-syllable words. As a first step, we explain how we obtained fixed-length pitch vectors.

4.1 Fixed-length pitch vectors

To implement a linear mapping within the DLM framework, given *n* words, we need an $n \times p$ matrix *C* to represent words' pitch contours, and an $n \times q$ matrix *S* for words' meanings. Consider the form matrix *C*, and recall that the tokens in our dataset have unequal numbers of pitch measurement points because tokens with longer durations contain more measurement points. Furthermore, the raw data include missing values due to gaps in the pitch contours. However, the row vectors of *C* need to have the same fixed length *p*. To achieve this, we used GAMs to obtain pitch contours represented by p = 100-dimensional vectors in normalized time. There are several ways in which such fixed-length vectors can be generated, of which we explored three.

- **Method I** The first method fitted separate GAMs to the f0 contours of each of the individual tokens, i.e., 4283 independent gam models, and then extracted the predicted contours. This method generates pitch contours that stay as close as possible to the empirical pitch measurements. However, this method inevitably includes by-token measurement noise in the estimation of the contours. In the case of simple univariate linear regression, the predicted value for a data point (on the regression line) will deviate from the observed value for that data point; taking the observed data point as gold standard is at odds with statistical modeling. Similarly, for the present dataset of time-series of measurements, the observed curves are not the given gold standard. There are several sources of noise: articulatory stochastic noise in the articulation, noise in the audio recordings, and noise in the pitch measurements. Method I incorporates the combined noise originating from these sources. Therefore, Method I serves as a baseline that we expect to yield the least precise results. Methods II and III implement two ways of reducing this by-observed pitch contour measure noise.
- **Method II** The second method fitted a GAM to the f0 contours of all the tokens of words with a given tone pattern, extracting the smooth for time and the word-specific smooth, and combining these to obtain word-type-specific smooths. This method abstracts away from the influences of contextual factors on the realization of pitch. The resulting pitch vectors are identical for all the tokens of a given word type. We anticipated that this would be the optimal situation for learning, as within-type variation is eliminated. This method also has a theoretical motivation, namely, that it is unlikely that the contextualized embeddings generated by an AI model will capture the full richness of the thought of human speakers engaged in real, 30-minute long conversations.
- **Method III** The third method, following Chuang et al. (2024), obtains token-specific pitch vectors predicted by GAMs with all contextual factors included. For our data, we used the four GAMs fitted to the four tonal contexts, as reported above in Section 3. This method has the advantage, compared to Method I, of removing by-observation noise. Furthermore, compared to Method

II, it has the advantage of having pitch vectors that vary from token to token. Thus, this method is optimal for detecting the extent to which by-token semantics and by-token phonetics are aligned. The more the contextualized embeddings diverge from the true semantic intentions of the speakers, the less well this method will perform.

After obtaining the estimated pitch vectors from the GAMs, we applied by-token normalization by centering and scaling each predicted pitch vector. By doing so, the mapping from meaning to form is forced to learn to predict the shape of pitch contours rather than the absolute pitch values of each token, which vary substantially across word types and speakers.

4.2 Contextualized embeddings

We made use of Contextualized Embeddings (CEs) to represent words' meanings. Word embeddings (semantic vectors) represent meanings in a distributed manner, building on the hypothesis that similar words occur in similar contexts (Harris, 1954; Landauer and Dumais, 1997; Mikolov et al., 2013). The first generation of semantic embeddings, such as fastText (Bojanowski et al., 2017), is fully determined by words' orthographic forms. However, a single orthographic form can express different meanings (e.g., English 'bank'), or different senses (e.g., Mandarin % (*shui3ping2*, 'level or horizontal position' or 'skill or proficiency')). Typically, the context in which a word is used provides disambiguating information. Contextualized Embeddings (CEs) were developed to provide token-specific, context-sensitive embeddings that capture the subtle differences in what a word may actually mean in context.

The CEs used in the current study were derived from a pre-trained unidirectional language model based on the GPT-2 architecture, developed by CKIP, Academia Sinica. Each token in our dataset was assigned a 768-dimensional vector representing its contextualized embedding.

To inspect the quality of the contextualized embedding space, we reduced the 768-dimensional semantic space to two dimensions using t-SNE (Van der Maaten and Hinton, 2008). Figure 8 displays embeddings in the resulting 2-D plane, with convex hulls highlighting clusters of tokens corresponding to different word types. Tokens clearly cluster by word, as expected. Furthermore, some semantically related words have clusters that are close to each other. For instance, in the middle-right of the Figure, the tokens of 大學 (*da4xue2*, 'university'), 學校 (*xue2xiao4*, 'school'), 國中 (*guo2zhong1*, 'middle school'), and 高中 (*gao1zhong1*, 'high school') occur closely together, which makes sense as they are all semantically related to educational institutions. Other school-related words such as 學 \pm (*xue2sheng1*, 'students') and 老師 (*lao3shi1*, 'teacher') also appear near these words. Some word clusters contain outliers. For instance, in the future'), and 後來 (*hou4lai2*, 'afterwards') has an outlier near 之後 (*zhi1hou4*, 'after').

4.3 Method

Modeling was conducted using the same dataset that we used above for the word-type-based analysis (see Table 2), which contains 4,283 tokens. This dataset, comprising all four tonal contexts, was split into a training dataset (80.39%) and a testing dataset (19.61%). Every word type was represented in both the training and testing data, with tokens per word being split roughly proportionally with 80% in the training dataset and 20% in the testing dataset.

Using the training data, we computed a linear mapping G from a 3443 × 768 semantic matrix S to a 3443 × 100 form matrix C by solving SG = C (for technical details, see Gahl and Baayen, 2024; Heitmeier et al., 2025). We then evaluated the quality of the mapping for both the training and the testing dataset.

The accuracy of a predicted pitch vector \hat{c} was evaluated as follows. For a given \hat{c} , we calculated its Euclidean distance to all gold-standard pitch vectors in C. We then identified its closest form



Figure 8: Contextualized embeddings, obtained from a pre-trained Chinese GPT-2 model, are shown in a two-dimensional plane obtained with t-SNE. Convex hulls (grey polygons) highlight the clusters of word types.

neighbor of \hat{c} . If this nearest neighbor was a token of the same word type as the target token, the predicted form vector was assessed as correct; otherwise, it was considered incorrect.

4.4 Results

We estimated three linear mappings from the same semantic matrix S with CEs to three different form matrices C, one for each of the three kinds of smoothed pitch contours introduced above.

The mean accuracy of method I was 2.8% on the training dataset and 1.4% on the testing dataset. The mean accuracy of method II was 23.5% on the training dataset and 15.1% on the testing dataset. The mean accuracy of method III was 12.3% on the training dataset and 7.7% on the testing dataset. All accuracies were above a permutation baseline of 0.4% and a majority baseline 1.3%, albeit by only a small margin for method I. That method I has the lowest accuracy is unsurprising, fitting GAMs to individual pitch contours unavoidably comes with overfitting and a loss of generalizability. Methods II and III gain strength from other tokens and incorporate less by-item observation noise. For these two methods, the mapping from meaning to pitch contours is substantially more accurate than would be expected under chance conditions. The best-performing method is method II, which abstracts away from the influences of contextual factors on the realization of pitch. This suggests that some abstraction away from the immediate context is helpful, possibly because the contextualized embeddings are not precise enough. After all, these embeddings come from a general large language model trained on large volumes of data that most likely diverge considerably for the language experience of the speakers interviewed for the Corpus of Spoken Taiwan Mandarin.

The results obtained with especially methods II and III clarify that there appears to be considerable isomorphy between the contextualized embedding space and the pitch space of word tokens. This isomorphy implies that if we take the most typical embedding for a given tone pattern and map it into the pitch space, the resulting predicted pitch contours should closely resemble the pitch contours identified by the GAM for that tone pattern. Figure 9 shows that this prediction is on the right track. The pitch contours shown in black are the contours predicted by the GAMs for the different tone patterns. They represent the best de-noised estimates of the average tone-pattern-specific pitch contours, and serve as our gold-standard pitch contours. These GAM-based pitch contours were obtained by first extracting the tone-pattern specific pitch contours for each of the four tonal contexts, and then averaging these. (These GAM-based contours were shown above in red in Figure 5 before). An alternative method, that results in nearly indistinguishable pitch contours, combines the data for all four contexts, and adds the tone-pattern specific contours to the general contour for female speakers.

We now consider how well these average pitch contours can be predicted from words' contextualized embeddings. The most typical embedding for a given tone pattern can be approximated by calculating the centroid of the contextualized embeddings of the tokens with this particular tone pattern. The centroid is simply the mean of the semantic vectors. When we think of embeddings as points in a high-dimensional space, the centroid is located at the center of these points. To obtain the centroid of a given tone pattern, we first obtained the centroid of every relevant word type by averaging the CEs of its tokens. We then obtained the centroid of the tone pattern by averaging the centroids of the word types associated with that tone pattern. In this way, every word type is given equal weight when determining the centroids for the tone patterns.

In order to get a sense of the semantics represented by these centroid vectors, we calculated, for each tone pattern, which contextualized embeddings are closest to the corresponding centroids. Table 5 lists, for each tone pattern, the two word types with embeddings that are closest to the centroids. These two word types provide an indication of the prototypical meaning of a tone pattern. For example, 她們 (*ta1men0*, 'they, female') and 他們 (*ta1men0*, 'they, male') are the most prototypical word types for T1-T0 tone pattern. The tone pattern T2-T0 appears to have 兒子(*er2zi0*, 'son') and 孩子 (*hai2zi0*, 'child') as prototypical members. Most tone patterns, however, are characterized by function words.

To obtain the pitch contours predicted for the tone patterns, we provided the centroids of the

Table 5: The top two word types which have contextualized embeddings that are closest to the centroid embedding of the 20 tone patterns. Proximity is evaluated using Euclidean distance.

Tone pattern	Top one closest word type	Top two closest word type
T1-T0	她們 (talmen0, 'they')	他們 (talmen0, 'they')
T1-T1	一些 (yilxiel, 'some')	一般 (yilban1, 'general')
T1-T2	當然 (dang1ran2, 'of course')	之前 (zhilqian2, 'before')
T1-T3	剛好 (gang1hao3, 'just right')	一起 (yi1qi3, 'together')
T1-T4	之後 (zhi1hou4, 'afterwards')	之類 (<i>zhillei4</i> , 'and so on')
T2-T0	兒子 (er2zi0, 'son')	孩子 (hai2zi0, 'child')
T2-T1	人家 (ren2jia1, 'others')	國中 (guo2zhong1, 'middle school')
T2-T2	其實 (qi2shi2, 'actually')	別人 (bie2ren0, 'others')
T2-T3	還有 (hai2you3, 'also')	還好 (hai2hao3, 'it's okay')
T2-T4	然後 (ran2hou4, 'then')	一樣 (yi2yang4, 'the same')
T3-T0	你們 (<i>ni3men0</i> , 'you all')	我們 (wo3men0, 'we')
T3-T1	很多 (hen3duo1, 'many')	女生 (nv3sheng1, 'girls')
T3-T2	起來 (qi3lai2, 'get up')	以前 (yi3qian2, 'before')
T3-T3	只有 (zhi3you3, 'only')	有點 (you3dian3, 'a bit')
T3-T4	以後 (yi3hou4, 'afterwards')	好像 (hao3xiang4, 'seems like')
T4-T0	這麼 (<i>zhe4me0</i> , 'so')	那麼 (na4me0, 'that')
T4-T1	那些 (na4xie1, 'those')	那邊 (na4bian1, 'over there')
T4-T2	個人 (ge4ren2, 'individual')	不然 (bu4ran2, 'otherwise')
T4-T3	到底 (dao4di3, 'after all')	那裡 (na4li3, 'there')
T4-T4	算是 (suan4shi4, 'considered as')	上面 (shang4mian4, 'above')

20 tone patterns as input to the three linear mappings defined above. The resulting predicted pitch contours are shown in Figure 9. Each panel in this trellis graph presents the estimates for a given tone pattern. The gold-standard pitch contours (obtained with our GAM models as described above) are presented in black, and the contours predicted by the three DLM mappings are color-coded. The contours obtained with method I are shown in blue, those obtained with method II in green, and those obtained with method III in red. For all three methods, the resulting predicted contours are similar, and often remarkably similar, to the gold-standard contours.



Figure 9: Pitch contours of 20 tone patterns in selected four tonal contexts. The black curves present pitch contours identified by GAM, estimated by the partial effect for tone_pattern, combined with a general contours of time female speakers (shown as the red curves in Figure 5, which is reproduced here). The blue, green, and red curves represent pitch contours predicted from the centroid of the contextualized embeddings using the three methods, respectively. For the blue curves, form vectors were obtained with GAM smooths fitted to individual word tokens. For the green curves, form vectors were obtained from a GAM fitted to the tokens of all words with a given tone pattern. For the red curves, form vectors were obtained with GAM smooths that included all contextual factors.

To assess the similarity between the gold-standard pitch contours and the pitch contours predicted using the meaning-to-pitch mappings, we first calculated the cosine similarity, averaged across the 20 tone patterns, between the gold-standard contours and each of the three DLM pitch contours. The contours from method II show a closer fit (cosine similarity 0.81) compared to the contours from method I (cosine similarity 0.59) and III (cosine similarity 0.66). The mean correlation between GAM-predicted pitch contours and three DLM-derived contours is 0.69, 0.82, and 0.78, respectively. However, when using Euclidean distance to evaluate similarity, method II scores slightly worse than the other two (1.48, 1.57, and 1.43, respectively). Figure 10 presents the distributions of these three

measures for the three methods, using boxplots. Regardless of how exactly the precision of the predictions is evaluated at the level of centroids, the three methods appear to perform with comparable accuracy, with a slight advantage for Method II when evaluated with the correlation and cosine similarity measures.



Figure 10: Boxplot showing the correlation, cosine similarity, and Euclidean distance between GAM-predicted pitch contours and DLM-derived pitch contours across 20 tone patterns.

4.5 Summary

In this section, we have shown that a simple linear mapping can predict the realization of tokenspecific pitch contours from its token-specific meaning in context with above-chance accuracy. This finding extends the earlier results of Chuang et al. (2024), which focused on disyllabic words with one specific tone pattern only (the rise-fall tone pattern T2-T4). Mapping accuracy for all 20 tone patterns is unsurprisingly somewhat lower that that observed by Chuang et al. (2024) for the T2-T4 tone pattern (30%–40% for training data and 27%–35% for testing data). Nevertheless, even for the present more varied dataset, accuracy is substantially above the majority baseline. This result is surprising in the light of the measurement noise that is inevitably present in both our pitch measurements and in the contextualized embeddings. The contextualized embeddings represent the knowledge of an artificial intelligence, trained on vast amounts of texts. The embeddings it generates must diverge from the meanings that the individual speakers had in mind. Nevertheless, the contextualized embeddings are sufficiently precise to enable far above chance prediction accuracy for word tokens' pitch contours. Interestingly, the 20 canonical tone patterns are surprisingly well approximated by projecting the centroids of the contextualized embeddings of the words with these tone patterns into the f0 space. In other words, the average pitch contours identified by the GAMs correspond to average contextualized embeddings in semantic space.

5 General discussion

The current study investigated the realization of pitch contours of disyllabic words in a corpus of spontaneous spoken Taiwan Mandarin. We first made use of the Generalized Additive Models to decompose f0 contours into a series of nonlinear functions of normalized time, with each function

representing the way in which a predictor modulates the pitch contour over time. A range of predictors was taken into account, including normalized time, gender, tonal context, tone pattern, speech rate, word position, bigram probability, and speaker. Surprisingly, the GAMs provided strong support for word-specific modulations of the pitch contours. Replacing word by word sense further improved model fit, which suggests that the effect of word may be semantic in nature. If so, the theory of the Discriminative Lexicon Model predicts that it should be possible to well approximate the token-specific pitch contours observed in the corpus with predicted token-specific pitch contours obtained with mappings taking the contextualized embeddings of the words in the corpus as input, and producing the corresponding pitch contours as output. We found that indeed a mapping from GPT-2 generated contextualized embeddings to 100-dimensional fixed-length pitch vectors predicts words' pitch contours with accuracies that are far above a majority choice baseline. Thus, our study successfully extends the meaning-to-pitch mapping from the T2-T4 tone pattern studied by (Chuang et al., 2024) to all tone patterns in Taiwan Mandarin. Our study also dovetails well with the evidence for the importance of word and sense as co-determinants of pitch contours reported by Lu et al. (2024) and Jin et al. (2024).

A remarkable finding is that words and their meanings co-determine the realization of the f0 contours in disyllabic words with effect sizes that considerably exceed those of tone pattern. This finding for disyllabic words aligns with previous research on Mandarin monosyllabic words (Jin et al., 2024), which reported that while the effect of tone pattern on pitch contours is modest, the effect of word is substantial. For disyllabic words, the stronger effect of word largely overshadows the effect of tone pattern.

Our results suggest that there are not only remarkable similarities, but also some clear differences, between tonal realization in laboratory speech and tonal realization in the Corpus of Spoken Taiwan Mandarin. Xu (1997) analyzed the pitch contours of 16 bi-tonal combinations using the /mama/ sequence. Among these combinations, only *ma1ma1* corresponds to a real word in Mandarin, (*ma1ma1*, "mother"); all the others are nonsensical combinations that are unnatural for native speakers. In their study, the f0 contours were carefully controlled, accounting for factors such as gender and speaking rate. Although laboratory speech and spontaneous speech differ in many ways, it is still informative to compare the two registers. We therefore reproduced Figure 3 from Xu (1997) (blue curves) and overlaid it with the DLM-derived f0 contours from Figure 9 (orange curves). In Figure 11, the pitch contours from Xu (1997) and our predicted contours are remarkably similar for several tone patterns, including T4-T4, T2-T3, and T2-T4. However, some tone patterns, such as T1-T1 and T1-T2, exhibit noticeable differences in pitch contours. These differences can be attributed to dialect differences, differences between spontaneous speech and laboratory speech, and differences between meaningful and meaningless words.

At this point, it might be objected that in this approach to Mandarin tone, it is unclear how tone sandhi could be accounted for. How would it be possible that, if indeed every word has its own pitch contour, all words with the T3-T3 tone pattern undergo the same phonological process, such that they become indistinguishable from words with the T2-T3 tone pattern? Our answer to this question has an empirical part and a theoretical part.

On the empirical side, in conversational Taiwan Mandarin, the two tone patterns are basically identical. For instance, in Figure 9, the tone patterns for T2-T3 and T3-T3 are quite similar. A detailed study of this tone sandhi (Lu et al., 2024) supports complete neutralization for Taiwan Mandarin. In other words, the words with the T3-T3 tone pattern can simply be re-classified as words with the T2-T3 tone pattern. There is no need to call on a rule of tone sandhi. In fact, even for standard Mandarin, as gauged by Xu (1997), the differences between T3-T3 and T2-T3 are hardly visible to the eye. However, T3-T3 tone sandhi has been reported to be incomplete for standard Mandarin (Yuan and Chen, 2014).

This brings us to the theoretical aspect of the question raised above, namely, how to account for tone sandhi processes in general. Within the framework of the Discriminative Lexicon model, as a model for highly automatized lexical processing, it is impossible to derive forms from forms,



Figure 11: The f0 contours for 16 tone patterns from the current study, based on the Corpus of Spoken Taiwan Mandarin, are compared with f0 contours from a previous study (Xu, 1997) on carefully controlled laboratory speech. The blue curves represent the f0 contours for 16 tone patterns from the controlled laboratory speech, reproduced from Figure 3 in Xu (1997). The orange curves correspond to the three LDL-predicted pitch contours from Method I, II, and III, as shown in Figure 9, and are reproduced here. These f0 contours are overlaid for comparison. Since the neutral tone (T0) was not included in Xu (1997), only 16 bi-tonal combinations are presented here.

a method widely used in educational settings. Forms are predicted from meanings. Importantly, Figures 9 and 11 show that the tone contours associated with tone patterns emerge straightforwardly from the meaning-to-form mapping, without the model ever being informed about tone patterns. In other words, a 'word and paradigm' approach (Blevins, 2016) to tonal realization appears to be quite feasible.

A question for further research is how to interpret the present findings for tone patterns with the neutral tone. As can be seen in Figure 9, for three of the four tone patterns, the overall pitch contour appears to be an almost linearly descending pitch contour. This could be viewed as another instance of tone sandhi in classical phonology, whereas within the DLM, this patterning would follow from words' contextual meanings. We leave this question for further research.

The results obtained in the present study have several theoretical implications. First, we have documented that the mapping from context-sensitive meaning to pitch contours is machine-learnable. It remains an open question whether human learners also generate pitch contours from semantics. The finding that just a linear mapping (from a statistical prespective, a straightforward multivariate multiple regression model) is all that is needed suggests that human speakers should also be able to learn this simple mapping between meaning and form. Importantly, our results are based on patterns of usage in a corpus of unscripted spontaneous speech, and the mere existence of these patterns indicates that language users must be absorbing community norms for tonal realization, albeit most likely subliminally.

Second, our findings challenge the axioms of the arbitrariness of the sign and the dual articulation of language. If the relation between form and meaning would be truly and fundamentally arbitrary, this would imply learning words and their meanings is extremely difficult, and would not allow any generalization. All that can be done is learn by heart that a form x is associated with a meaning y. However, our simple linear mapping generalizes to held-out data. This falsifies the axiom that the relation between form and meaning (here, pitch and meaning) is completely arbitrary.

Third, preliminary results, reported in Chuang et al. (2023), suggest that English two-syllable words with left stress also have pitch contours that have strong word-specific pitch components. The study by Schmitz et al. (2025) reports similar findings for English three-constituent compounds. The accumulating evidence poses a new challenge for linguistics: understanding why these isomorphies between form and meaning exist, irrespective of whether a language is a tone language or a stress language.

Funding

This work was supported by the European Research Council under Grant SUBLIMINAL (#101054902) awarded to R. Harald Baayen.

Acknowledgements

The authors thank Yu-Hsiang Tseng for identifying word sense for the dataset in the current manuscript.

Appendix 1: dataset and model summary

Tone Pattern	Tonal Context				
	4.4	3.4	4.1	4.0	
T1-T0	4	4	4	2	
T1-T1	15	15	12	12	
T1-T2	18	15	17	11	
T1-T3	8	8	6	6	
T1-T4	23	19	18	22	
T2-T0	7	6	6	4	
T2-T1	11	9	9	11	
T2-T2	14	9	11	8	
T2-T3	14	12	13	12	
T2-T4	23	19	22	12	
Т3-Т0	4	3	4	4	
T3-T1	10	8	6	7	
T3-T2	16	12	11	12	
T3-T3	11	10	9	8	
T3-T4	18	15	17	11	
T4-T0	3	4	3	1	
T4-T1	10	9	12	11	
T4-T2	24	20	17	15	
T4-T3	9	8	8	5	
T4-T4	46	35	45	36	
Total	288	240	250	210	

Table A.1: Overview of the four sub-datasets grouped by the four tonal contexts, with the number of word types for each tone pattern.

A generativity of affinity	Estimate	Ctd Emman	4	
A. parametric coefficients	Estimate	Sta. Error	t-value	p-value
(Intercept)	5.2965	0.0251	210.6574	< 0.0001
gendermale	-0.5283	0.0343	-15.3930	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(normalized_t):genderfemale	1.0020	1.0025	24.4870	< 0.0001
s(normalized_t):gendermale	2.6783	2.9183	7.6161	< 0.0001
s(speaker)	51.0328	53.0000	72.0201	< 0.0001
s(normalized_t,word)	1832.2415	2592.0000	6.9980	< 0.0001
s(normalized_t,tone_pattern)	111.9037	179.0000	2.3357	< 0.0001
s(speech_rate):genderfemale	2.1305	2.5343	3.8114	0.0126
s(speech_rate):gendermale	2.3418	2.7025	9.7111	< 0.0001
ti(normalized_t,speech_rate)	2.9792	3.0132	30.9735	< 0.0001
s(norm_utt_pos)	1.7416	2.1134	101.1452	< 0.0001
ti(normalized_t,norm_utt_pos)	6.6304	7.9364	6.7134	< 0.0001
s(bg_prob_prev)	2.8803	2.9825	32.2276	< 0.0001
ti(normalized_t,bg_prob_prev)	4.9995	6.2443	2.3813	0.0252
s(bg_prob_fol)	1.0072	1.0139	4.8743	0.0265
ti(normalized_t,bg_prob_fol)	7.8925	8.6541	9.2356	< 0.0001

Table A.2: Summary of the model fitted with word for the 4.4 tonal context dataset

Table A.3: Summary of the model fitted with word for the 3.4 tonal context dataset

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	5.3009	0.0239	222.0659	< 0.0001
gendermale	-0.5151	0.0309	-16.6894	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(normalized_t):genderfemale	2.8018	2.9407	11.3801	< 0.0001
s(normalized_t):gendermale	1.0026	1.0035	0.6147	0.4328
s(speaker)	49.9064	54.0000	24.2216	< 0.0001
s(normalized_t,word)	1430.9843	2160.0000	5.8479	< 0.0001
s(normalized_t,tone_pattern)	92.8198	179.0000	1.3530	< 0.0001
s(speech_rate):genderfemale	2.3721	2.7269	9.3361	< 0.0001
s(speech_rate):gendermale	2.1719	2.5746	3.4047	0.0351
ti(normalized_t,speech_rate)	6.7124	7.9532	4.9905	< 0.0001
s(norm_utt_pos)	1.0004	1.0009	144.2261	< 0.0001
ti(normalized_t,norm_utt_pos)	6.4372	7.6506	4.8472	< 0.0001
s(bg_prob_prev)	2.4885	2.7657	15.9136	< 0.0001
ti(normalized_t,bg_prob_prev)	8.5179	8.8649	26.3242	< 0.0001
s(bg_prob_fol)	1.0023	1.0044	8.8528	0.0029
ti(normalized_t,bg_prob_fol)	2.3852	2.7236	3.6728	0.0093

Estimate	Std. Error	t-value	p-value
5.2709	0.0251	210.1867	< 0.0001
-0.4851	0.0331	-14.6522	< 0.0001
edf	Ref.df	F-value	p-value
1.0003	1.0004	14.3424	0.0002
2.4742	2.7866	3.8965	0.0197
50.3703	54.0000	28.1287	< 0.0001
1520.9285	2250.0000	5.5411	< 0.0001
88.4589	179.0000	1.2859	< 0.0001
1.0005	1.0010	5.3846	0.0203
1.9702	2.3439	2.9261	0.0393
6.2226	7.4529	14.0239	< 0.0001
1.0006	1.0011	28.5619	< 0.0001
7.3974	8.3940	2.4661	0.0054
2.0687	2.4374	12.4622	< 0.0001
3.6914	4.7848	2.8590	0.0170
2.4158	2.7373	2.3562	0.0475
7.8947	8.6462	7.8430	< 0.0001
	Estimate 5.2709 -0.4851 edf 1.0003 2.4742 50.3703 1520.9285 88.4589 1.0005 1.9702 6.2226 1.0006 7.3974 2.0687 3.6914 2.4158 7.8947	EstimateStd. Error5.27090.0251-0.48510.0331edfRef.df1.00031.00042.47422.786650.370354.00001520.92852250.000088.4589179.00001.00051.00101.97022.34396.22267.45291.00061.00117.39748.39402.06872.43743.69144.78482.41582.73737.89478.6462	EstimateStd. Errort-value5.27090.0251210.1867-0.48510.0331-14.6522edfRef.dfF-value1.00031.000414.34242.47422.78663.896550.370354.000028.12871520.92852250.00005.541188.4589179.00001.28591.00051.00105.38461.97022.34392.92616.22267.452914.02391.00061.001128.56197.39748.39402.46612.06872.437412.46223.69144.78482.85902.41582.73732.35627.89478.64627.8430

Table A.4: Summary of the model fitted with word for the 4.1 tonal context dataset

Table A.5: Summary of the model fitted with word for the 4.0 tonal context dataset

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	5.2497	0.0312	168.0817	< 0.0001
gendermale	-0.4725	0.0407	-11.6043	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(normalized_t):genderfemale	2.9192	2.9806	29.4234	< 0.0001
s(normalized_t):gendermale	1.0003	1.0005	5.0343	0.0248
s(speaker)	49.0454	52.0000	30.1430	< 0.0001
s(normalized_t,word)	1251.5978	1890.0000	6.4034	< 0.0001
s(normalized_t,tone_pattern)	93.3188	179.0000	1.4397	< 0.0001
s(speech_rate):genderfemale	2.6175	2.8809	8.3676	0.0001
s(speech_rate):gendermale	2.7772	2.9508	8.2388	0.0001
ti(normalized_t,speech_rate)	6.8744	7.9913	6.6015	< 0.0001
s(norm_utt_pos)	2.0931	2.4575	41.5639	< 0.0001
ti(normalized_t,norm_utt_pos)	7.0265	8.0821	12.4450	< 0.0001
s(bg_prob_prev)	1.9451	2.3101	12.5304	< 0.0001
ti(normalized_t,bg_prob_prev)	5.9749	7.1845	4.1940	0.0001
s(bg_prob_fol)	1.0006	1.0010	8.6352	0.0033
ti(normalized_t,bg_prob_fol)	5.6756	6.9167	3.1157	0.0029

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	5.2990	0.0255	207.7527	< 0.0001
gendermale	-0.5276	0.0345	-15.2859	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(normalized_t):genderfemale	1.0037	1.0046	22.3822	< 0.0001
s(normalized_t):gendermale	2.6128	2.8854	7.3128	0.0001
s(speaker)	50.7591	53.0000	59.8792	< 0.0001
s(normalized_t,sense_type)	1645.3569	2394.0000	6.1036	< 0.0001
s(normalized_t,tone_pattern)	115.1964	179.0000	2.3487	< 0.0001
s(speech_rate):genderfemale	2.2619	2.6469	3.5374	0.0310
s(speech_rate):gendermale	2.2986	2.6760	8.0901	0.0001
ti(normalized_t,speech_rate)	3.9985	4.7382	15.5029	< 0.0001
s(norm_utt_pos)	2.3278	2.6858	57.3866	< 0.0001
ti(normalized_t,norm_utt_pos)	5.6890	7.1236	4.8475	< 0.0001
s(bg_prob_prev)	2.8814	2.9813	24.2073	< 0.0001
ti(normalized_t,bg_prob_prev)	5.3450	6.5906	2.7958	0.0083
s(bg_prob_fol)	1.1588	1.2896	7.5202	0.0028
ti(normalized_t,bg_prob_fol)	7.8741	8.6372	11.9355	< 0.0001

Table A.6: Summary of the model with sense_type for the 4.4 tonal context dataset

Table A.7: Summary of the model with sense_type for the 3.4 tonal context dataset

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	5.3190	0.0245	217.3049	< 0.0001
gendermale	-0.5279	0.0322	-16.4049	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(normalized_t):genderfemale	2.7782	2.9300	11.8593	< 0.0001
s(normalized_t):gendermale	1.0016	1.0022	0.7333	0.3918
s(speaker)	49.3383	53.0000	23.1424	< 0.0001
s(normalized_t,sense_type)	1311.3647	1980.0000	5.6701	< 0.0001
s(normalized_t,tone_pattern)	90.0576	179.0000	1.2569	< 0.0001
s(speech_rate):genderfemale	2.8229	2.9667	15.2549	< 0.0001
s(speech_rate):gendermale	2.7810	2.9558	3.5426	0.0186
ti(normalized_t,speech_rate)	7.5150	8.4879	7.7827	< 0.0001
s(norm_utt_pos)	2.5743	2.8458	47.1183	< 0.0001
ti(normalized_t,norm_utt_pos)	7.0027	8.0972	5.2953	< 0.0001
s(bg_prob_prev)	2.4559	2.7397	11.7554	< 0.0001
ti(normalized_t,bg_prob_prev)	8.5396	8.8648	32.3655	< 0.0001
s(bg_prob_fol)	1.0011	1.0022	8.6211	0.0033
ti(normalized_t,bg_prob_fol)	2.3724	2.7114	3.1644	0.0175

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	5.2785	0.0258	204.6860	< 0.0001
gendermale	-0.4733	0.0330	-14.3633	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(normalized_t):genderfemale	1.0006	1.0008	10.6364	0.0011
s(normalized_t):gendermale	2.2744	2.6385	4.7004	0.0076
s(speaker)	49.3758	53.0000	25.2692	< 0.0001
s(normalized_t,sense_type)	1390.6999	2052.0000	5.2771	< 0.0001
s(normalized_t,tone_pattern)	89.6240	179.0000	1.2507	< 0.0001
s(speech_rate):genderfemale	1.0012	1.0023	0.8458	0.3574
s(speech_rate):gendermale	1.0900	1.1655	8.0534	0.0043
ti(normalized_t,speech_rate)	6.2987	7.3658	11.8930	< 0.0001
s(norm_utt_pos)	1.0007	1.0014	20.7997	< 0.0001
ti(normalized_t,norm_utt_pos)	4.4585	5.9266	1.7014	0.1187
s(bg_prob_prev)	2.3038	2.6419	4.8417	0.0042
ti(normalized_t,bg_prob_prev)	5.4972	6.7456	3.7973	0.0005
s(bg_prob_fol)	2.6706	2.8966	4.2319	0.0049
ti(normalized_t,bg_prob_fol)	7.3670	8.3279	3.9638	0.0001

Table A.8: Summary of the model with sense_type for the 4.1 tonal context dataset

Table A.9: Summary of the model with sense_type for the 4.0 tonal context dataset

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	5.2524	0.0327	160.7046	< 0.0001
gendermale	-0.4743	0.0414	-11.4693	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
s(normalized_t):genderfemale	2.9426	2.9856	31.6209	< 0.0001
s(normalized_t):gendermale	1.0016	1.0023	7.1126	0.0076
s(speaker)	48.0016	53.0000	23.1946	< 0.0001
s(normalized_t,sense_type)	994.9116	1539.0000	5.5151	< 0.0001
s(normalized_t,tone_pattern)	85.9056	179.0000	1.1669	< 0.0001
s(speech_rate):genderfemale	2.0016	2.3892	2.3245	0.1149
s(speech_rate):gendermale	2.1820	2.5488	4.7352	0.0061
ti(normalized_t,speech_rate)	7.1264	8.2127	7.6463	< 0.0001
s(norm_utt_pos)	1.8211	2.1735	46.3419	< 0.0001
ti(normalized_t,norm_utt_pos)	6.9596	8.0649	10.2243	< 0.0001
s(bg_prob_prev)	2.1932	2.5430	13.3047	< 0.0001
ti(normalized_t,bg_prob_prev)	6.8339	7.8695	6.2328	< 0.0001
s(bg_prob_fol)	1.0194	1.0344	7.0325	0.0073
ti(normalized_t,bg_prob_fol)	1.9840	2.6243	0.4195	0.7027

Appendix 2: the effects of segments and frequency

Our findings demonstrate that the word itself is a strong predictor of pitch contours in disyllabic words. However, one might question whether this robust effect is at least partially influenced by segmental properties, given existing evidence on the impact of vowel height and onset consonants on Mandarin tones (Ho, 1976a; Ladd and Silverman, 1984; Ohala and Eukel, 1976; Whalen and Levitt, 1995). Additionally, lexical frequency has long been recognized as a factor influencing f0 contours, with lower-frequency words being produced with higher pitch (Zhao and Jurafsky, 2007). To address these concerns, this additional analysis clarifies the effects of word's segmental composition and lexical frequency on pitch contours.

Following Chuang et al. (2024), for our disyllabic words, we coded four predictors for segments. vowel1_height and vowel2_height are the vowel height of the first syllable and the second syllable, respectively. Each has five levels: 'high', mid', and low', low-high' and mid-high'. onset1_type and onset2_type are the type of the onset consonant of the first syllable and the second syllable, respectively. Each has seven levels: 'aspirated-affricate', 'aspirated-stop', 'unaspiratedaffricate', 'unaspirated-stop', 'voiceless-fricative', 'voiced', and 'null'. frequency represents the log-transformed count of occurrences of a word type in entire spoken corpus of Taiwan Mandarin.

We built up a baseline model that includes gender, tonal context, tone pattern, speaking rate, speaker, word position, bigram probability, but excludes word. To simplify the analysis, this model was based on an omnibus dataset that integrates all four tonal contexts.

To examine the effects of segments, four factor smooth terms for vowel1_height, vowel2_height, vowel1_type, and vowel2_type were added to the baseline model together.

```
baseline + s(normalized_t, vowel1_height, bs='fs', m=1)+
s(normalized_t, vowel2_height, bs='fs', m=1)+
s(normalized_t, onset1_type, bs='fs', m=1)+
s(normalized_t, onset2_type, bs='fs', m=1)
```

To examine the effect of frequency, we added the smooth term for frequency, in combination with its interaction with normalized_t, to the baseline model.

```
baseline + s(frequency, k=4)+
ti(normalized_t, frequency, k=c(4,4))
```

Table A.10 presents the improvement of model fit compared with the baseline model, as evaluated by AIC change. The inclusion of the four segmental predictors combined improves the model fit by 5,267.08 units, while the inclusion of word leads to a more substantial improvement of 21,532.92 units. Moreover, in Chuang et al. (2024), with only around 50 word types, the segment-related controls were highly confounded with one another, as indicated by the high concurvity scores of around 0.75. However, in our dataset that contains a greatly larger number of word types, these effects can be better disentangled. For the baseline + four segmental predictors model, the concurvity scores are much lower than those reported in Chuang et al. (2024): 0.42 for s (normalized_t, vowel1 _height), 0.40 for s (normalized_t, vowel2_height), 0.35 for s (normalized_t, onset1_type), and 0.35 for s (normalized_t, onset2_type). In the baseline + word model, the concurvity score for word is also low (0.12). Additionally, although baseline + four segmental predictors + word has the lowest AIC, concurvity of four segmental predictors is all 1, and that of word is 0.21, suggesting that segmental predictors are highly colinear with word when they are both present.

Similarly, the inclusion of word (21532.92 AIC units) resulted in a more substantial AIC decrease than frequency (205.76 AIC units). Besides, we added frequency on top of the baseline + word model, which resulted in a further AIC decrease of 10.01 units. However, in this model, concurvity was very high for frequency (0.99) and low for word (0.13).

Model	AIC	AIC difference	
baseline	279.70	-637646.74	-
baseline + frequency	291.03	-637852.51	-205.76
baseline + four segmental predictors	447.63	-642913.82	-5267.08
baseline + word	2475.79	-659179.66	-21532.92
<pre>baseline + frequency + word</pre>	2475.54	-659189.68	-21542.93
baseline + four segmental predictors	2444.84	-659198.24	-21551.50

Table A.10: Improvement of model	fit gauge	d by AIC	change
----------------------------------	-----------	----------	--------

Overall, word by itself contributes more to the model fit than the four segmental predictors combined, as well as lexical frequency. In line with Chuang et al. (2024), the effect of word cannot be simply reduced to the effect of segments. Besides, word is a better predictor of pitch contours than lexical frequency.

References

- Arnold, D. and Tomaschek, F. (2016). The Karl Eberhards corpus of spontaneously spoken southern german in dialogues — audio and articulatory recordings. In Draxler, C. and Kleber, F., editors, *Tagungsband der 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, pages 10–13, Muenchen. Ludwig-Maximilians-Universitaet.
- Arnon, I. and Priva, U. C. (2014). Time and again: The changing effect of word and multiword frequency on phonetic duration for highly frequent sequences. *The Mental Lexicon*, 9(3):377–400.
- Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. (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*.
- Blevins, J. P. (2016). Word and paradigm morphology. Oxford University Press.
- Boersma, P. and Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.0. 37, 2018.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Chang, C.-Y. (2010). *Dialect differences in the production and perception of Mandarin Chinese tones*. PhD thesis, The Ohio State University.
- Chao, Y. (1968). A Grammar of Spoken Chinese. University of California Press.
- Chuang, Y.-Y., Baayen, R. H., and Bell, M. J. (2023). Do words sing their own tunes? word-specific pitch realizations in Mandarin and English. In *Proceedings of ICPhS 2023*.
- Chuang, Y.-Y., Bell, M. J., Tseng, Y.-H., and Baayen, R. H. (2024). Word-specific tonal realizations in Mandarin. *arXiv preprint arXiv:2405.07006*.
- Chung, K. S. (2006). Contraction and backgrounding in Taiwan Mandarin. *Concentric: Studies in Linguistics*, 32(1):69–88.
- De Saussure, F. (1966). Course in General Linguistics. McGraw, New York.
- Drager, K. K. (2011). Sociophonetic variation and the lemma. *Journal of Phonetics*, 39(4):694–707.
- Ernestus, M. (2000). Voice assimilation and segment reduction in casual Dutch. A corpusbased study of the phonology-phonetics interface. LOT, Utrecht.
- Fon, J. (2004). A preliminary construction of Taiwan Southern Min spontaneous speech corpus. Technical report, Tech. Rep. NSC-92-2411-H-003-050, National Science Council, Taiwan.
- Fon, J. and Chiang, W.-Y. (1999). What does Chao have to say about tones?—a case study of Taiwan Mandarin. *Journal of Chinese Linguistics*, pages 13–37.

- Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, 84(3):474–496. Publisher: Linguistic Society of America.
- Gahl, S. and Baayen, R. H. (2024). Time and thyme again: Connecting English spoken word duration to models of the mental lexicon. *Language*, 100(4):623–670.
- Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.
- Harris, Z. S. (1954). Distributional structure. Word, 10(2-3):146–162.
- Heitmeier, M., Chuang, Y.-Y., and Baayen, R. H. (2025). *The Discriminative Lexicon: Theory and implementation in the Julia package JudiLing.* Cambridge University Press, Cambridge. in press.
- Ho, A. T. (1976a). The acoustic variation of Mandarin tones. *Phonetica*, 33(5):353–367.
- Ho, A. T. (1976b). Mandarin tones in relation to sentence intonation and grammatical structure. *Journal of Chinese Linguistics*, pages 1–13.
- Hsieh, P.-j. (2013). Prosodic markings of semantic predictability in Taiwan Mandarin. In *INTERSPEECH*, pages 553–557.
- Hsieh, S.-K., Tseng, Y.-H., Chou, H.-Y., Yang, C.-W., and Chang, Y.-Y. (2024). Resolving Regular Polysemy in Named Entities. arXiv:2401.09758 [cs].
- Huang, C.-R., Hsieh, S.-K., Hong, J.-F., Chen, Y.-Z., Su, I.-L., Chen, Y.-X., and Huang, S.-W. (2010). Chinese wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese information processing*, 24(2):14–23.
- Huang, K. (2018). Phonological identity of the neutral-tone syllables in Taiwan Mandarin: An acoustic study. *Acta Linguistica Asiatica*, 8(2):9–50.
- Jescheniak, J. D. and Levelt, W. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 20(4):824–843.
- Jin, X., Ernestus, M., and Baayen, R. H. (2024). A corpus-based investigation of pitch contours of monosyllabic words in conversational Taiwan Mandarin. *arXiv preprint arXiv:2409.07891*.
- Johnson, K. (2004). Massive reduction in conversational American English. In Spontaneous speech: data and analysis. Proceedings of the 1st session of the 10th international symposium, pages 29–54, Tokyo, Japan. The National International Institute for Japanese Language.
- Ladd, R. and Silverman, K. E. (1984). Vowel intrinsic pitch in connected speech. *Phonetica*, 41(1):31–40.

- Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.
- Lohmann, A. (2018). Cut (n) and cut (v) are not homophones: Lemma frequency affects the duration of noun–verb conversion pairs. *Journal of Linguistics*, 54(4):753–777.
- Lu, Y., Chuang, Y.-Y., and Baayen, R. H. (2024). Form and meaning co-determine the realization of tone in Taiwan Mandarin spontaneous speech: the case of Tone 3 sandhi. *arXiv preprint arXiv:2408.15747*.
- Martinet, A. (1965). *La Linguistique Synchronique: Études et Recherches*. Presses Universitaires de France, Paris.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Ohala, J. J. and Eukel, B. W. (1976). Explaining the intrinsic pitch of vowels. *The Journal* of the Acoustical Society of America, 60(S1):S44–S44.
- Pitt, M., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.
- Plag, I., Homann, J., and Kunter, G. (2017). Homophony and morphology: The acoustics of word-final s in English. *Journal of Linguistics*, 53(1):181–216.
- Plag, I., Lohmann, A., Hedia, S. B., and Zimmermann, J. (2020). An <s>is an <s'>, or is it? plural and genitive-plural are not homophonous. *Complex words: Advances in morphology*, page 260.
- Saito, M. (2024). Enhancement effects of frequency: An explanation from the perspective of Discriminative Learning. Doctoral dissertation, University of Tübingen.
- Saito, M., Tomaschek, F., Sun, C.-C., and Baayen, R. H. (2024). Articulatory effects of frequency modulated by semantics. *Interfaces of Phonetics*, 38:125.
- Schmitz, D. (2022). Production, perception, and comprehension of subphonemic detail: Word-Final /s/ in English. Language Science Press.
- Schmitz, D., Plag, I., and Bell, M. J. (2025). Modeling the relationship between prominence and semantics in English compounds. Paper presented at the Workshop on Morphological Variation, 47th Annual Meeting of the German Linguistic Society (DGfS), Mainz, Germany.
- Shih, C. (2000). A declination model of Mandarin Chinese. In *Intonation: Analysis, modelling and technology*, pages 243–268. Springer.
- Stanford, J. N. (2016). Sociotonetics using connected speech: A study of Sui tone variation in free-speech style. *Asia-Pacific Language Variation*, 2(1):48–82.

- Tang, K. and Bennett, R. (2018). Contextual predictability influences word and morpheme duration in a morphologically complex language (kaqchikel mayan). *The Journal of the Acoustical Society of America*, 144(2):997–1017.
- Team, R. C. (2020). R Core Team R: a language and environment for statistical computing. *Foundation for Statistical Computing.*
- Tomaschek, F., Plag, I., Ernestus, M., and Baayen, R. H. (2021). Phonetic effects of morphology and context: Modeling the duration of word-final s in English with naïve discriminative learning. *Journal of Linguistics*, 57(1):123–161.
- Turnbull, R. (2017). The role of predictability in intonational variability. *Language and speech*, 60(1):123–153.
- Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).
- Wang, S.-F. (2024). Contrast and predictability in the variability of tonal realizations in Taiwan Southern Min. In *Proc. SpeechProsody* 2024, pages 542–546.
- Whalen, D. H. and Levitt, A. G. (1995). The universality of intrinsic f0 of vowels. *Journal* of phonetics, 23(3):349–366.
- Wood, S. N. (2017). Generalized additive models: an introduction with R. CRC press.
- Wu, Y., Adda-Decker, M., and Lamel, L. (2020). Mandarin Lexical Tones: A Corpus-Based Study of Word Length, Syllable Position and Prosodic Position on Duration. In *Interspeech 2020*, pages 1908–1912, Shanghai, China. ISCA.
- Wu, Y., Lamel, L., and Adda-Decker, M. (2021). Tone realization in Mandarin speech: a large corpus based study of disyllabic words. In 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), pages 1–5. IEEE.
- Xu, C. (2024). Cross-dialectal perspectives on Mandarin neutral tone. *Journal of Phonetics*, 106:101341.
- Xu, C. X. and Xu, Y. (2003). Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association*, 33(2):165–181.
- Xu, Y. (1997). Contextual tonal variations in Mandarin. Journal of phonetics, 25(1):61–83.
- Xu, Y. and Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *The Journal of the Acoustical Society of America*, 111(3):1399–1413.
- Yuan, J. and Chen, Y. (2014). 3rd tone sandhi in standard Chinese: A corpus approach. *Journal of Chinese Linguistics*, 42(1):218–237.
- Zhao, L. (2023). *Production and perception of lexical tone variation in Mandarin dialects*. PhD thesis, University of York.
- Zhao, Y. and Jurafsky, D. (2007). The effect of lexical frequency on tone production. In Proceedings of the 16th International Congress of Phonetic Sciences, pages 477–480. International Phonetic Association.

Zimmermann, J., Carignan, C., and Tyler, M. D. (2016). Morphological status and acoustic realization: Findings from New Zealand English. In *Proceedings of the 16th australasian international conference on speech science and technology*, pages 6–9.