TONE AS A FACTOR INFLUENCING THE DYNAMICS OF DIPHTHONG REALIZATIONS IN STANDARD MANDARIN

Chenyu Li[1], Jalal Al-Tamimi[1], Yaru Wu[2,3,4]

[1]Université Paris Cité, LLF, CNRS; [2]CRISCO, Université de Caen Normandie; [3]LPP, CNRS, Sorbonne Nouvelle; [4]LISN, CNRS, Université Paris-Saclay

chenyu.li@etu.u-paris.fr; jalal.al-tamimi@u-paris.fr; yaru.wu@unicaen.fr

ABSTRACT

This paper investigates the impact of the tonal factor on the realization of diphthongs in Standard Mandarin. We employed a set of Generalized Additive Mixed Models (GAMMs) to respectively test whether and how tones and fundamental frequency ($f0$) influence the realization of the diphthong /ai/ in a journalistic speech corpus of Standard Mandarin. Results show, in general, the realizations of /ai/ to significantly differ with respect to tones, due to different heights and contours of $f0$. A high tone results in a more closed and more front /ai/ and a low tone results in a more open and more back realization; a rising tone strengthens the dynamic features of /ai/ while a falling tone weakens such features and leads to monophthongization. The results suggest that the relation between $f0$ and vowel realization, found in monophthongs, is equally applicable to diphthongs: $f0$ is negatively correlated with F1 and positively with F2.

Keywords: Standard Mandarin, diphthong, tone, $f0$, Generalized Additive Mixed Models

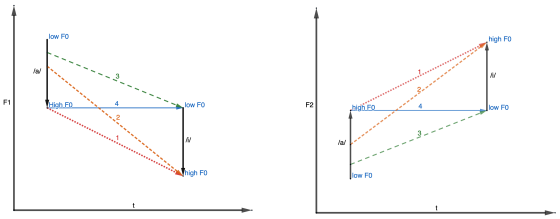
1. INTRODUCTION

Phonological segments and suprasegmental features are two major areas of research in phonology and phonetics, but they are often studied separately. The tone, as a suprasegmental concept, is a major component of Mandarin phonology [1] that was widely studied in monosyllabic sequences [2] or in continuous speech [3]. In Standard Mandarin, there are four lexical tones (with Chao digits): tone 1 (high tone - 55), tone 2 (rising tone - 35), tone 3 (low tone - 21 in natural speech) and tone 4 (falling tone - 51) [1]. In many languages, like Mandarin, there is a dynamic vowel concept, the diphthong, which is considered as a continuously varying single segment [4, 5], or a dynamic vowel sequence which has two targets and a transition between them [6].

The interaction between tones (contours of $f0$) and single vowels has been extensively studied, with accumulating evidence of this interaction in the forms of intrinsic $f0$ and the tonal effect on vocalic realization. A large body of research have confirmed that the intrinsic $f0$ is attested in many languages, including tonal, such as Mandarin [7], and non-tonal languages [8, 9]. Results suggest that a more closed vowel will have a higher intrinsic $f0$. Regarding the effect of $f0$ on vowel realization, different studies have evaluated the impact of different $f0$ levels on vowel realization from an acoustic [7, 10] and an articulatory view point [11, 12, 13]. In general, a higher $f0$ will cause the vowel to be realized as more closed and more front. However, until now, there does not exist any clear evidence for the interaction between the $f0$ and diphthong realization.

Although there is evidence from perception [14] suggesting that this interaction could also apply to the case of diphthongs, there has been no research showing direct evidence of it. As a matter of fact, most of the current literature on Mandarin diphthongs is at the descriptive [6] or the historical/dialectal (e.g., [15]) level and does not take $f0$ into account.

The present study aims to explore how tones influence the realizations of diphthongs in Standard Mandarin and to assess whether the effect of the $f0$ on vowel realization occurs in diphthongs. We hypothesize that for /ai/, the vowel realizations will be affected by $f0$; more concretely, a higher $f0$ will make the vowel produced as more closed and more front, whereas a lower $f0$ will make the vowel produced as more open and more back. Under this hypothesis, we can make various predictions for how /ai/ will be realized with different tones.



(a) F1 **(b)** F2

Figure 1: Predictions following hypothesis

According to Fig. 1, the F1 curve of /ai/ will be

higher with a low falling tone (tone 3) and lower with a high tone (tone 1), flatter with a falling tone (tone 4) and steeper with a rising tone (tone 2). The F2 curve will have an opposite pattern to that of F1.

2. METHODS

2.1. Corpus and data processing

The data came from a journalistic speech corpus from the LDC [16], while only a 2h long data set was provided by the LISN lab to this study for processing and further analyses. The data were automatically segmented, aligned and transcribed with Chinese characters and their phonetic transcriptions with lexical tones at the word and the phoneme levels, using a speech transcription system from LISN in forced alignment mode [17]. Only lexical tones (1 to 4) in Standard Mandarin are included in the lexicon. Four falling diphthongs, i.e., a diphthong whose first component is syllabic [6]: /ai/, /au/, /əi/ and /əu/ were processed [18]. We present the results of /ai/ here due to it being the most common diphthong in the database.

Segmental boundaries between the diphthong /ai/, the previous and following segments were manually verified using *Praat* (version 6.1.54) [19], with the start and end boundaries being chosen using the positions of the *onset* and *offset* of F2 of /ai/, following [20]. We considered the duration of the tone as being that of the *rime*, i.e., from the *onset* of the vowel to *offset* of the syllable [21]: both diphthong and tone had the same duration. This yielded a data set of 215 occurrences: 112 for 8 female speakers (9 for tone 1, 18 for tone 2, 15 for tone 3 and 70 for tone 4) and 103 for 15 male speakers (4 for tone 1, 26 for tone 2, 17 for tone 3 and 56 for tone 4). For each occurrence, we have the following information: the tone; the position of /ai/ in the word, i.e., final or non-final position and monosyllabic; the previous and following segments; speaker's gender and the anonymized identity.

The acoustic data were extracted for each occurrence, with the fundamental frequency ($f0$), and the formant frequencies for F1 and F2. Formant frequencies were automatically obtained using the *Burg* method (window length = 25ms, pre-emphasis from 30 Hz, maximum frequency = 5500 Hz for female and 5000 Hz for male, and a maximum of 5 formants). We then obtained $f0$ measurements using the two-passes method, with auto-correlation (following [22], implemented in [23]); $f0$ estimations are speaker-dependent based on their range that prevents errors in extraction. For each occurrence, we obtained 11 time normalized intervals, at 10% interval for formant and $f0$ frequencies. This yielded a total of 2301 data points (1105 for male speakers, 1196 for female speakers).

2.2. Modelling

Since the diphthong /ai/ is a dynamic vowel sequence, we used Generalized Additive Mixed Models (GAMMs) [24] to capture its dynamic pattern. GAMMs are a statistical technique that models the non-linear relation between predictors and an outcome varying in the time domain. Modelling with GAMMs was done using the packages *mgcv* [24] (version 1.8-33, for modelling) and *itsadug* [25] (version 2.4, for visualization) in R computing language (version 4.2.2) [26]. For model specification, we followed recommendations from [27, 28, 29].

For our specific dataset, we performed two sets of modelling. First, to evaluate the interaction between the diphthong realization using F1 and F2 dimensions and the tonal unit, our predictor was the `tone` (categorical - four levels) and the outcome either F1 or F2 frequencies. We used `time` as a continuous predictor (11 normalized intervals), represented via a *smooth* as a non-linear variable, to track the dynamic pattern during the diphthong realization. The other variables, i.e., the speaker ID, the consonant at syllable onset and the position in the word, were considered as factor smooths modelled as random effects. Different genders were modelled separately to assess differences between the two genders and to reduce computational power.

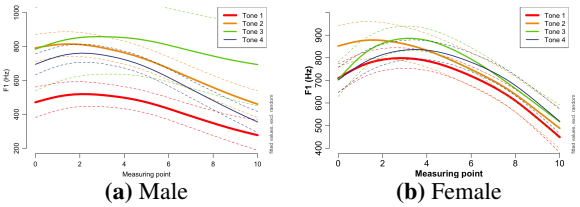
We used an *ordered* categorical predictor for `tone` to reduce the Type I error and increase power in our model [29]. We used a *by* interaction term to adjust for the variable levels of tones for our smooth terms. Then, we verified the auto-correlation levels of our model and obtained an Auto-Regressive GAMMs. After fitting the model, we performed a `gam.check` and the data suffered from a "long tail" distribution; this was corrected using the `"scat"` family [27, 29, 30]. The comparison of the models with `gam.check` showed that the modified ones fitted the data better.

The second model has roughly the same structure, except that the categorical predictor `tone` is replaced by two continuous ones: `f0` and `duration`, to assess the correlation between tone and $f0$/duration, because tone, as a phonological unit, is represented phonetically by the $f0$ contour over time [1].

3. RESULTS

3.1. First model

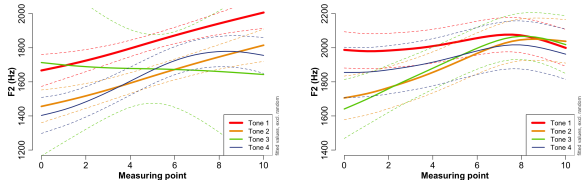
The purpose of the first model is to evaluate any possible influences of tone on the diphthong realization. Figs. 2 and 3 illustrate the predictions obtained from the model for F1 and F2 curves, differentiated by the four tones.



(a) Male **(b)** Female

Figure 2: F1 curves with 4 tones

All F1 contours are falling from interval 2 to 8. In tone 1 (high level), the curve is generally lower than in other tones, signalling a more closed realization of the diphthong /ai/. The curve with tone 3 (low falling), is higher than in other tones, which indicates that /ai/ would be realized as the most open. The curves with tone 1 and 3 have a relatively flat pattern, which indicates partial monophthongization. The curve with tone 2 (rising) starts relatively high (800 Hz) and stops relatively low (600 Hz), which indicates minimal monophthongization, compared with the other tones. The curve with tone 4 (falling) of the female speakers (Fig. 2b) starts relatively low (700 Hz) and ends relatively high (700 Hz), showing a flat pattern which is related to the monophthongization. Using the plot_diff function in the *itsadug* package, we report on the regions of statistical differences (95% of CI). We found that the curves of male speakers are significantly different in the 1-2, 1-3, 1-4, 2-4 (first and final parts), 3-4 (final part) tonal pairs; 1-2 (first part), 1-3, 1-4, 2-3 (starting part), 2-4 (first part) pairs for the female ones are significantly different.



(a) Male **(b)** Female

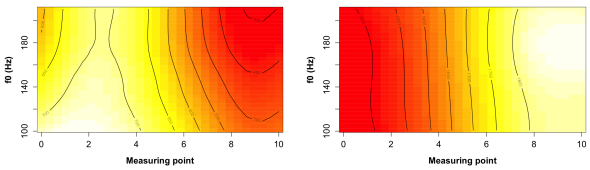
Figure 3: F2 curves with 4 tones

As for the F2 (Fig. 3), the results point to rising contours over time, except for tone 3 in male speakers (Fig. 3a). The curve with tone 1 is generally higher than in other tones, i.e., /ai/ with tone 1 would be realized as most front. The curve with tone 3 of female speakers is generally lower than those with the other tones. For the male speakers, although the curve with tone 3 does not follow the rising trend, it is lower than those with the other tones in the second part (intervals 6 - 10). This indicates that /ai/ would be realized as most back with tone 3. We observed a pattern with a steeper slop in the curve with tone 2, indicating a minimal monophthongization. Based on the predictions of our hypothesis (Fig. 1), tone 4 (falling), is predicted to monophthongize the diphthong. Fig. 3 reveals a flatter curve with tone 4, on the full contour for female speakers and between intervals 5-10 for male speakers. The significance test (95% of CI) show that the curves of pairs 1-2, 1-4 for male speakers, and 1-2, 1-3, 1-4 (first part), 2-4 (starting part), 3-4 (starting part) pairs for female speakers, are significantly different from each other.

The predictions of the first model generally confirm the hypothesis: the tonal effect on vowel realization is observed on the diphthong /ai/ in this study. More concretely, with a high tone (tone 1), /ai/ tends to be realized as more closed and more front; with a low tone (tone 3), it is more open and more back; with a rising tone (ton 2), it tends to have a typical diphthongized realization; and with a falling tone (tone 4), the diphthong tends to be monophthongized (but see *Summary and discussion* for specific issues related to tone 4).

3.2. Second model

In the second model, we consider the combination of $f0$ and duration, to more accurately model the impact of tones on the realization of the diphthong /ai/. To visualize the result of the second model, we selected three intervals at the quartiles, i.e., 25%, 50% and 75% of the duration. We present the summed effects 3D plots using the function fvisgam. In the figures, the dependent variable, i.e., F1 or F2 for female or male speakers, is determined by two different continuous predictors on two axes, i.e., the normalized time on the x axis and the $f0$ on the y axis, with F1 or F2 changes on the z axis, whereby the lighter the colour, the higher the frequency values of F1 or F2; and vice versa.

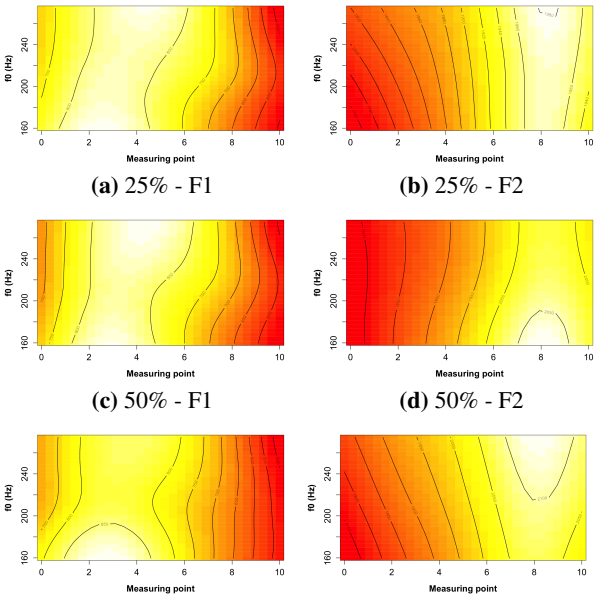


(a) F1 **(b)** F2

Figure 4: F1/F2 with different $f0$ (Male, 50% of duration)

The Fig. 4 demonstrates the model's predictions for male speakers at 50% of the duration due to minimal differences across all three portions (at 25-50-75%); this specific portion shows that the effect of $f0$ on F1 or F2 is independent from the effect of duration on F1 or F2. Overall, F1 falls during the realization of the diphthong, whereas F2 rises. Meanwhile, the contour lines (which show the same value of F1 or F2) and colors show that

when $f0$ rises, F1 falls and F2 rises (especially in the transition part).



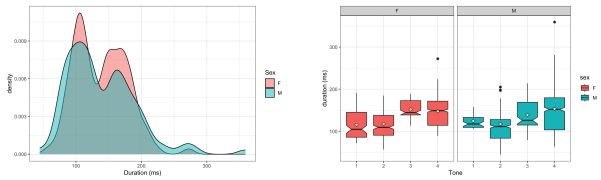
(a) 25% - F1 **(b)** 25% - F2

(c) 50% - F1 **(d)** 50% - F2

(e) 75% - F1 **(f)** 75% - F2

Figure 5: F1/F2 with different $f0$ (Female)

The patterns of the female speakers are more complex. According to Fig. 5, we found that the predictions at 50% of the duration (Figs. 5c and 5d) suggest that F1 has a positive correlation with $f0$ and F2 has a negative correlation (from interval 4); results that are contra to the predictions of our hypothesis, which is likely due to the fact that the reliability of model predictions in some specific cases is influenced by the amount of data. Fig. 6a shows that the distribution of duration of female speakers is *bi-modal* than that of male speakers, which makes the predictions obtained at 50% of duration less reliable and a more dynamic account is needed.



(a) Duration density of genders **(b)** Duration of each tone

Figure 6: Duration of /ai/

4. SUMMARY AND DISCUSSION

In this paper, we were interested in modelling how the diphthong /ai/ in standard Mandarin is realized and assessed any possible relationships with tone, on the one hand, and with the combination of $f0$ and duration as multi-dimensional predictors, on the other. The first model predicts well how the /ai/ is realized in different tones. Overall, if it is associated with a high tone (tone 1), /ai/ is produced as more closed and more front. If it is associated with a low tone (tone 3), the diphthong is produced as more open and more back. /ai/ would be realized as more diphthongized when it is associated with a rising tone (tone 2). As for a falling tone (tone 4), the expected monophthongized pattern has only been confirmed in the result of F1 and F2 curve of female speakers. The second model allows for more refined modelling to capture the effects of $f0$ and of the duration on F1 or F2 and its change over time. In general, $f0$ is negatively correlated with F1, and positively with F2. The models and their predictions in this paper show a good fit to the data and allowed to evaluate how diphthongs are realized in Standard Mandarin. They also allowed to verify the hypotheses with respect to effects of tones and $f0$, independently of each other.

The curves with tone 4, according to the hypothesis, should be more flat, but from the prediction obtained from the model, they show an obvious dynamic pattern. In Figs. 2a and 3a, the F1 and F2 curves with tone 4 for male speakers show a "diphthongized" dynamic feature similar to that with tone 2, which contradicts our hypothesis. From Fig. 6b, the duration of /ai/ varies with different tones: in tones 4 for female and male speakers and in tone 3 for female speakers, the duration is the longest. This allows for more time to complete the vowel realization from the *onset* to the *offset*. In the visualizations (Figs. 2 and 3), the "compression" effect of normalized time will cause curves with longer absolute duration to appear more dynamic. In this case, the duration effect is possibly more important than the $f0$ effect. The results of the model 2 and of the significance test in model 1 also highlighted the impact of gender on F1 or F2. The effects of male speakers are more reflected on F1 (aperture) level, while those for female speakers are more reflected on F2 (backness) level.

The present study has revealed that in addition to $f0$, duration and gender can also affect diphthong realizations. Further research may focus on two aspects. Firstly, through comparative studies of different dialects in Mandarin, we aim to investigate the universality of the tonal influence on diphthongs and provide explanations for various phenomena of monophthongization in different dialects. Secondly, we will attempt to incorporate the current study into Articulatory Phonology framework and explain its relationship with the intrinsic $f0$ of vowels.

5. ACKNOWLEDGEMENTS

This work was supported by the Laboratoire de Linguistique Formelle, the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083), contributing to the IdEx Université Paris Cité - ANR-18-IDEX-0001. The data set was supported by LISN, as part of the IdEx Emergence project *Son-Discours*. Thanks to the *Chinese Student Council* for funding. Special thanks to Prof. Ioana Chitoran.

6. REFERENCES

[1] S. Duanmu, *The phonology of standard Chinese*. OUP Oxford, 2007.

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

[3] Y. Xu, F. Liu *et al.*, "Tonal alignment, syllable structure and coarticulation: Toward an integrated model," *Italian Journal of Linguistics*, vol. 18, no. 1, p. 125, 2006.

[4] B. Malmberg, *Structural linguistics and human communication: An introduction into the mechanism of language and the methodology of linguistics*. Springer Science & Business Media, 2012, vol. 2.

[5] P. Delattre, "Comparing the vocalic features of English, German, Spanish and French," 1964.

[6] H. Ren, "On the acoustic structure of diphthongal syllables," Ph.D. dissertation, University of California, Los Angeles, 1986.

[7] P. Wang, *A Statistical Study on the Tones and Vowels of Beijing Dialect*. PhD Thesis, Nankai University, 2007.

[8] D. H. Whalen and A. G. Levitt, "The universality of intrinsic f0 of vowels," *Journal of phonetics*, vol. 23, no. 3, pp. 349–366, 1995.

[9] J.-M. Hombert, "Consonant types, vowel height and tone in Yoruba," *UCLA Working Papers in Phonetics*, vol. 33, pp. 40–54, 1976.

[10] D. Erickson, R. Iwata, M. Endo, and A. Fujino, "Effect of tone height on jaw and tongue articulation in Mandarin Chinese," in *International symposium on tonal aspects of languages: With emphasis on tone languages*, 2004.

[11] F. Hu, "Tonal effect on vowel articulation in a tone language," in *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, 2004.

[12] P. Hoole and F. Hu, "Tone-vowel interaction in standard Chinese," in *International symposium on tonal aspects of languages: With emphasis on tone languages*, 2004.

[13] J. A. Shaw, W.-r. Chen, M. I. Proctor, and D. Derrick, "Influences of tone on vowel articulation in Mandarin Chinese," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 6, pp. S1566–S1574, 2016.

[14] O. Niebuhr, "Intrinsic pitch in opening and closing diphthongs of German," in *Speech Prosody 2004, International Conference*, 2004.

[15] W. Zhang, *Evolution and competition: changes in the phonological structure of the Guanzhong dialect*. Peaple's Press of Shaanxi Province, 2002.

[16] A. Morris, B. Antonishek, X. Li, and S. Strassel, *HAVIC MED Progress Test–Videos, Metadata and Annotation*. Linguistic Data Consortium, University of Pennsylvania, 2019.

[17] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[18] C. Li, *The role of factors influencing the realization of diphthongs in Standard Mandarin: dynamic modeling with GAMMs*. Master Thesis, niversité Paris Cité, 2022.

[19] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[20] Ö. Ünal-Logacev, S. Fuchs, and L. Lancia, "A multimodal approach to the voicing contrast in Turkish: Evidence from simultaneous measures of acoustics, intraoral pressure and tongue palatal contacts," *Journal of Phonetics*, vol. 71, pp. 395–409, 2018.

[21] J. M. Howie and J. M. Howie, *Acoustical studies of Mandarin vowels and tones*. Cambridge University Press, 1976, vol. 18.

[22] J. Al-Tamimi and G. Khattab, "Acoustic correlates of the voicing contrast in Lebanese Arabic singleton and geminate stops," *Journal of Phonetics*, vol. 71, pp. 306–325, 2018.

[23] J. Al-Tamimi. (2022) Praat *f*0 Accurate Estimation. https://github.com/JalalAl-Tamimi/Praat-f0-Accurate-Estimation.

[24] S. N. Wood, *Generalized additive models: an introduction with R*. chapman and hall/CRC, 2006.

[25] J. Van Rij, M. Wieling, R. H. Baayen, and D. van Rijn, "itsadug: Interpreting time series and autocorrelated data using gamms," 2015.

[26] R Core Team, *R: A Language and Environment for Statistical Computing*, https://www.R-project.org/, R Foundation for Statistical Computing, Vienna, Austria, 2021.

[27] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English," *Journal of Phonetics*, vol. 70, pp. 86–116, 2018.

[28] M. Sóskuthy, "Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction," *arXiv preprint arXiv:1703.05339*, 2017.

[29] M. Soskuthy, "Evaluating generalised additive mixed modelling strategies for dynamic speech analysis," *Journal of Phonetics*, vol. 84, p. 101017, 2021.

[30] Y.-Y. Chuang, J. Fon, I. Papakyritsis, and H. Baayen, "Analyzing phonetic data with generalized additive mixed models," in *Manual of clinical phonetics*. Routledge, 2021, pp. 108–138.