



Apical vs. Regular Vowel Duration: A Corpus-based Analysis of Contextual Influences in Standard Mandarin

Jingyi Sun¹, Bowei Shao¹, Martine Adda-Decker^{1,2}

¹Laboratoire de Phonétique et Phonologie, CNRS & Sorbonne Nouvelle, France

²Laboratoire Interdisciplinaire des Sciences du Numérique, CNRS & Université Paris-Saclay, France

jingyi.sun@sorbonne-nouvelle.fr, bowei.shao@sorbonne-nouvelle.fr,
martine.adda-decker@sorbonne-nouvelle.fr

Abstract

The syllabic fricatives, traditionally known as apical vowels in Standard Mandarin, share the place of articulation with their sibilant onsets, exhibiting a narrow tongue tip constriction at the dental/alveolar or postalveolar region. We hypothesize that, like voiced fricatives, they face aerodynamic constraints where voicing and frication cannot be fully optimized simultaneously. The narrow constriction may delay intra-oral pressure release, leading to unstable and delayed voicing onset and consequently shorter acoustic durations. To test this hypothesis while minimizing segmentation errors, we applied forced alignment to a large-scale, style-comparable corpus to analyze their durational differences. The results show that in both spontaneous and read speech, the syllabic fricatives are significantly shorter compared to regular vocalic nuclei. The findings join previous articulatory and acoustic evidence and provide further evidence on the fricative nature of these syllabic segments.

Index Terms: syllabic fricatives, duration, forced alignment, Mandarin Chinese

1. Introduction

1.1. Phonetics and phonology of the syllabic fricatives in Standard Mandarin

The syllabic fricatives [ʒ, ʐ], traditionally known as apical vowels and noted as [ɿ] and [ʅ] [1, 2], are a complex set of nucleic segments in Chinese languages. In Standard Mandarin (SM), they are generally considered as allophones of the high front vowel /i/ [3, 4, 5]. The two syllabic fricatives have a co-occurrence restriction with the high front vowel /i/ following dental and retroflex sibilants, e.g., *[si] *[ʂi]. In these two contexts, in place of the high front vowel, there occurs [ʒ] and [ʐ], which only occur in these contexts respectively, e.g., [sʒ], [ʂʐ].

It has been shown that these syllabic fricatives' phonological behaviour mirrors that of the vowels: they are the nucleus of the syllable and constitute the tone-bearing unit of a syllable. They undergo various phonological processes targeting syllabic nucleus, such as contextual devoicing and tone sandhi [2].

Articulatorily, there is solid evidence showing that the syllabic fricatives are homorganic with their co-occurring onsets. In [sʒ] and [ʂʐ], for example, [s] and [ʒ], [ʂ] and [ʐ] share the same place of articulation [3, 2, 5, 6]. During the production of these syllables, there is virtually no tongue tip movement from the onsets [s, ʂ] to the nuclei [ʒ, ʐ] [7, 8].

Acoustically, some amount of frication noise may be observed during the first part of the syllabic fricatives superimposed on voicing. This frication noise is interpreted either as transitional frication noise introduced by the sibilant onsets [5], or as an inherent part of the nuclei [9]. The first interpretation

leads to the conclusion they may be syllabic approximants [ɹ, ɻ] [4, 5], based on the observation that frication noise is visible mainly in the beginning of the nuclei. A clear formant pattern emerges once the frication noise fades out. The second interpretation builds upon the fact that the tongue tip does not move during the transition from the onsets to the nuclei, the presence of frication noise, which can persist until the mid-portion of the nuclei, is thus additional evidence of the fricative nature of these segments, and argues for a voiced syllabic fricative analysis [2, 9].

These two seemingly antagonistic analyses are, in our view, describing the same phenomenon: trade-off between voicing and frication in the production of voiced fricatives [10, 11].

1.2. Aerodynamic voicing constraint in the production of voiced fricatives

The aerodynamic voicing constraint (AVC) has long been recognized [12, 13]. Ohala [10, 11] considers the AVC as one of the aerodynamic principles on which speech production relies.

To achieve optimal voicing during speech, the trans-glottal pressure difference between sub-glottal pressure ($P_{sub-glott}$) and oral cavity pressure (P_{oral}) should be maximized. In the case of fricatives, another aerodynamic requirement must be satisfied to produce frication: a large pressure drop across the oral constriction must be achieved. The atmospheric pressure ($P_{atmospheric}$) cannot be modulated by the speaker; therefore, one must try to keep P_{oral} as high as possible.

Thus, the aerodynamics of voiced fricatives faces an inherent conflict: for voicing, the P_{oral} should be as low as possible, whereas for frication, it should be as high as possible. Both demands cannot be satisfied optimally at the same time. This is shown as the pressure hierarchy for voiced fricatives in 1, which the syllabic fricatives in SM must obey. Strong voicing reduces fricative energy, creating an approximant-like shape. Strong frication, however, may compromise voicing [11].

$$P_{sub-glott} > P_{oral} > P_{atmospheric} \quad (1)$$

1.3. Hypothesis on the duration of syllabic fricatives

Our hypothesis is that the syllabic fricatives are acoustically shorter due to the complexity in maneuvering the aerodynamics compared to regular vocalic nuclei. This hypothesis is built on two empirical findings: i) the aerodynamic hierarchy necessary for the production of voiced fricatives, and ii) the homorganicity between the syllabic fricatives and their respective onsets.

Generally, in a regular syllable such as [sa], for the voicing of [a] to start, the aerodynamic and the articulatory maneuvers work in synergy. The lowering of the tongue body and the jaw for the articulation of the vowel [a] opens the oral cavity, caus-

ing a rapid P_{oral} drop which equalizes the pressure difference between P_{oral} and $P_{atmospheric}$ efficiently. The trans-glottis pressure difference is thus maximized and the voicing of the vowel starts.

In syllables containing syllabic fricatives such as [sz], the articulatory maneuver is largely absent as presented above. The oral cavity is connected to atmosphere only through a narrow air channel and a rapid P_{oral} drop would be impossible. A certain time would be necessary for P_{oral} to be low enough to allow voicing. These two scenarios are presented in Figure 1.

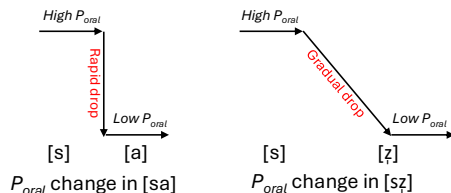


Figure 1: Schematic diagram of oral pressure drop from onset [s] to nuclei [a] and [z].

The gradual drop in P_{oral} acoustically results in delayed and unstable voicing onset of the nuclei [z, z̥], manifesting as shorter acoustic duration.

While pulse-level manual annotation can yield precise voicing onset labels [9], it is not feasible for large-scale analysis. To ensure consistency and avoid human-induced variability, we test this hypothesis using a large speech corpus annotated with the Montreal Forced Aligner (MFA) 3.0.6 [14], which applies uniform acoustic criteria for phoneme boundary detection.

2. Method

2.1. Corpus

To examine potential duration differences between syllabic fricative nuclei and vocalic nuclei, we used a style-comparable subset of a previously published Mandarin corpus [15], comprising spontaneous dialogues and read speech from 30 native speakers (17 males, 13 females). Each speaker pair engaged in a spontaneous conversation, recorded in stereo at 48 kHz/16 bit. Based on selected segments from their own spontaneous speech, each speaker later read a rewritten, formal version twice, recorded in mono. The study protocol was approved by the ethics board of Université Sorbonne Nouvelle (CER-USN) and validated by the CNRS Data Protection Officer (ref. 2-24092). The final dataset includes 20 h of spontaneous and 6 h of read speech.

To avoid confounding effects in duration measurements from medials, diphthongs, nasal codas, and isolated [z̥] syllables, we focused on CV or CC structures where consonants /ts^h s ts̥ s̥ z/ are followed by a single vowel /a, ə, u/ or a syllabic fricative /z, z̥/. To extract target vowel and syllable durations, we implemented an automated transcription and forced-alignment workflow, followed by manual verification:

1. Automatic transcription: The open-source Whisper [16]’s large model generated speech-to-text transcriptions with timestamped text files.
2. Initial forced alignment: We used MFA with the `spacy-pkuseg` package [17] for Chinese word segmentation. Its pretrained acoustic model treats the syllabic fricatives (/z/ and /z̥/) as independent phonemes and detects boundaries using MFCCs; spectrograms show nucleic onsets

for both [z, z̥] and regular vowels around the second or third pitch pulse.

3. Manual correction of transcriptions: We reviewed and refined Whisper’s textual output (e.g., misrecognized characters or word boundaries) in Praat 6.4.08 [18].
4. Final forced alignment: The corrected transcripts were then realigned with the original audio using MFA to obtain consistent phoneme-level segmentation.

Following this procedure, we obtained 46,477 syllables needed: 69.57% with [z, z̥] and 30.43% with [a, ə, u]; of which 81.41% came from spontaneous speech and 18.59% from read speech, with 36.35% contributed by female speakers and 63.65% by male speakers.

2.2. Statistical Modeling

We extracted syllable-internal durations and contextual variables using Python 3.9.19 with `tgt` and `pandas`. Two response variables—absolute vowel duration and its relative proportion in the syllable—were examined to test if apical vowels are systematically shorter than regular vowels.

Given the right-skewed duration distribution (most values below 0.3 s; KS statistic = 0.233, $p < 0.01$), we employed a Generalized Linear Mixed Model (GLMM) with a Gamma distribution and inverse link. In the GLMM, primary fixed effects (vowel type and vowel) compared regular vocalic nuclei versus syllabic fricative nuclei (further refined across five nuclei [a, ə, u, z, z̥]), while additional fixed effects included preceding consonant, tone, prosodic position (sentence-initial, medial, final, isolated), and speech style. Random intercepts for Chinese character and speaker controlled for lexical and speaker-specific variability, as characters sharing a syllable can differ in frequency, syntax, and prosody. We used the `bobyqa` optimizer to aid convergence.

Although vowel proportion was approximately normal, its 0–1 range warranted a Beta regression model with the same fixed and random effects, plus a natural spline (5 degrees of freedom) to capture nonlinearity between absolute and proportional durations.

Furthermore, since the verb “to be” (the character 是 [sz̥]) comprises 31.59% of tokens and has a significantly longer average duration (0.0982 s) than the overall mean (0.0831 s) of syllabic fricatives, we therefore introduced a binary variable (`is_shi`) to explicitly control for “to be” in our model, preventing an artificial inflation of differences between syllabic fricative nuclei and regular vocalic nuclei. All statistical analyses were conducted in R 4.4.2 using the `lme4` [19] and `glmmTMB` [20, 21] packages.

3. Results

3.1. Duration Comparison

We employed a GLMM with a Gamma distribution and an inverse link function to analyze the variability in absolute vowel duration (in seconds). Using vowel type (including syllabic fricatives and regular vowels) as a predictor variable, we found that regular vowels were significantly longer than syllabic fricatives by an average of 0.05s ($p < 0.001$). Additionally, Tone 2 (rising tone) significantly lengthened the duration of syllabic fricatives by 0.04s ($p < 0.001$). When considering the duration variability associated with the verb “to be” (the character 是 [sz̥]), the mean duration of syllabic fricatives increased significantly by 0.14s. Furthermore, vowels in sentence-initial and

sentence-medial positions exhibited significant shortening effects of 0.053s and 0.051s, respectively ($p < 0.001$).

Given that vowel duration still exhibited variation within the same vowel type, we conducted a finer-grained modeling by considering each vowel as a separate predictor variable. The baseline condition was set as follows: vowel /a/, preceding consonant /s/, isolated sentence, Tone 1, read speech, and the absence of the verb “to be”. Differences across characters and speakers were captured using random effects. The results showed that the variance at the character level ($\sigma^2 = 4.25$) was significantly greater than at the speaker level ($\sigma^2 = 1.77$), indicating that specific characters exerted a stronger influence on vowel duration than individual speaker habits. The influence of fixed factors on vowel duration is uneven and asymmetric. The

Table 1: Fixed Effects of the GLMM for Vowel on Duration

Cat.	Level	Coef (β)	SE	p	Δ (s)
Intercept	(Baseline)	3.88	0.62	***	0.26
Vowel					
	ə	+1.60	0.52	**	-0.07
	u	+2.08	0.46	***	-0.09
	z _r	+4.02	0.59	***	-0.13
	z _r	+3.52	0.52	***	-0.12
Consonant					
	ʃ	+0.32	0.54		-0.02
	ts	+0.21	0.53		-0.01
	tʃ	-0.29	0.51		+0.01
	ts ^h	+0.01	0.59		-0.00
	tʃ ^h	+1.17	0.53	*	-0.06
	z _r	-0.68	1.09		+0.03
Tone					
	2	-1.21	0.42	**	+0.04
	3	-0.20	0.50		+0.01
	4	+0.85	0.40	*	-0.02
Position					
	medial	+6.07	0.27	***	-0.16
	initial	+5.74	0.33	***	-0.15
	final	+0.45	0.28		-0.02
Style					
	Spon	+0.60	0.09	***	-0.02
Control					
	is_shi	-3.77	1.00	***	+0.10

Note: Cat. = Category; Level = Factor Level; Coef.(β) = Coefficient; SE = Standard Error; p = p-value; Δ (s) = Duration Change in seconds. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

following sections present how the interactions between fixed effects and vowel affect the duration of the nucleic segment. Table 1 summarizes the estimated coefficient, standard error, significance levels, and actual duration changes (converted to the original time scale using the inverse link function). The formula used for conversion is:

$$\text{Predicted Duration} = \frac{1}{\beta_0 + \sum \beta_i X_i} \quad (2)$$

As shown in Table 1, all vowels exhibited significant differences in duration compared to the baseline regular vowel /a/ ($p < 0.01$). When all conditions were averaged, /a/ had the longest duration, with a predicted value of 0.26s. Among all vowels, [z] showed the strongest shortening effect ($\Delta = 0.10$ s),

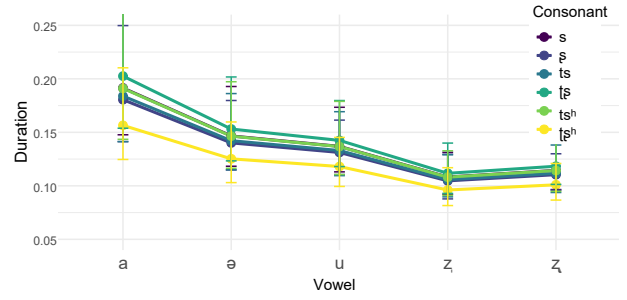


Figure 2: Interaction Between Vowel and Consonant on Vowel Duration

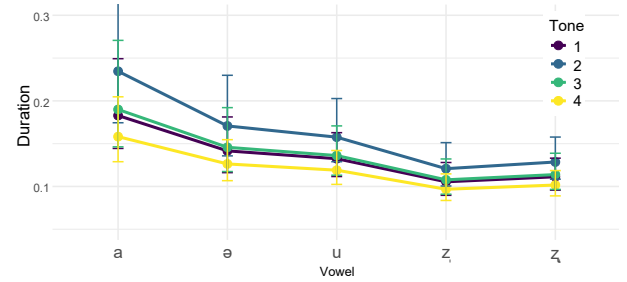


Figure 3: Interaction Between Vowel and Tone on Vowel Duration

followed by [z] ($\Delta = 0.11$ s), suggesting a positive correlation between the complexity of aerodynamic adjustment and duration compression in syllabic fricatives. Within regular vowels, /u/ exhibited a greater reduction in duration ($\Delta = 0.09$ s) than the schwa /ə/ ($\Delta = 0.07$ s).

3.1.1. Consonant Effect

Figure 2 illustrates that only the aspirated retroflex affricate /tʃ^h/ shortened the duration of the following vowel ($\Delta = 0.06$ s, $p = 0.028$), likely reflecting the erosive effect of its strong airflow on vowel steady-state duration. However, neither the aspirated alveolar affricate /ts^h/ nor other unaspirated consonants (e.g., /s/, /ts/) showed significant effects, indicating heterogeneity in consonant-vowel coarticulation. The five vowels exhibited a three-tiered hierarchical pattern in duration: /a/ remained the longest regardless of consonant pairing and showed greater variation, though its duration was significantly reduced when preceded by /tʃ^h/. /ə/ and /u/ were shorter than /a/, with /ə/ slightly longer than /u/. The two syllabic fricatives were the shortest, also with less duration variation. However, the retroflex [z_r] was slightly longer than the non-retroflex [z].

3.1.2. Tone Effect

Figure 3 indicates that Tone 2 (rising) lengthens vowel duration by +0.04 s ($p = 0.004$), especially for regular vowels like /a/, while Tone 4 (falling) shortens it by 0.02s ($p = 0.035$); Tone 3 shows no effect ($p = 0.686$). The influence of consonants and tones on syllabic fricatives was weaker than on regular vowels, suggesting that regular vowels are more sensitive to tonal modulation. These results are consistent with previous studies [22, 23, 24], which found that rising tones require a longer duration to accommodate their gradual f0 rise, whereas falling tones,

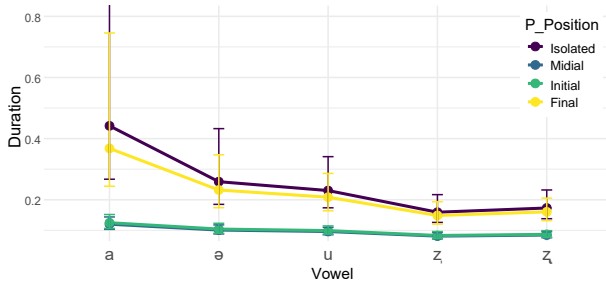


Figure 4: Interaction Between Vowel and Prosodic Position on Vowel Duration

with their steeper f0 drop, are produced more rapidly.

3.1.3. Prosodic Position Effect

In Figure 4, vowels in sentence-initial and sentence-medial positions exhibited the strongest shortening effects ($\Delta = 0.15s$, $p < 0.001$; $\Delta = 0.16s$, $p < 0.001$), whereas vowels in sentence-final positions did not significantly differ from those in isolated sentences ($\Delta = 0.02s$, $p = 0.28$). Additionally, regular vowels were more sensitive to prosodic position effects, showing greater variation in duration compared to syllabic fricatives.

3.1.4. Speech Style and High-Frequency Word Effect

Spontaneous speech exhibited an overall vowel duration reduction of 0.02s compared to read speech ($p < 0.001$). Moreover, when controlling for other fixed effects, speech style magnified duration differences among the five vowels: /a/ was significantly longer than all other vowels. The two syllabic fricatives were significantly shorter than regular vowels.

3.2. Relative Duration Proportion

We used a Beta-distributed GLMM with a logit link to model bounded responses between 0 and 1, using the same baseline conditions and random effects as in the vowel duration model.

Overall, the variation in relative vowel duration mirrors that of absolute duration, albeit with some complex associations. Specifically, under baseline fixed effects and without spline adjustments, vowels other than /a/ (9.2%) exerted negative influences on the proportion, with [ʒ] showing the strongest effect ($\Delta = 2.7\%$, $p < 0.001$). When the word is the verb “to be” 是, the vowel duration proportion increases by about 3.3%, suggesting a tendency for “to be” to exhibit a higher relative vowel duration. Aspirated consonants slightly reduced the proportion, while unaspirated affricates decreased it by 3.2%. Moreover, Tone 3 amplified the vowel duration proportion more than Tones 2 and 4. Although vowels in medial positions have shorter absolute durations, their relative proportions are 5% higher; vowels at the beginning and end of sentences increase by 2.5% and 3.3%, respectively. Finally, spontaneous speech shows a roughly 2% higher vowel duration proportion than read speech.

3.3. Nonlinear Effects of Vowel Duration and Proportion

Natural spline fitting results indicate a nonlinear relationship between absolute vowel duration and its relative proportion. To illustrate this, we present the distribution for duration < 0.3 s, covering 98% of the data. As shown in Figure 5, all vowels

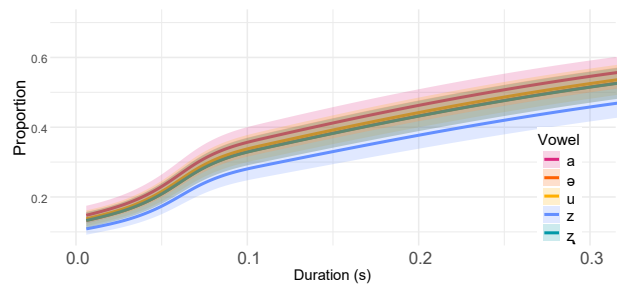


Figure 5: Nonlinear Relationship Between Vowel Duration and Relative Proportion (95% CI)

exhibit similar patterns in their internal syllabic duration distribution. When the absolute duration is < 0.07 s, the relative proportion of vowel duration accelerates with increasing absolute duration, and the differences among vowels remain minimal. However, beyond 0.07 s, the proportion increases linearly, and the differences among vowels become more pronounced.

Regarding individual vowels, /a/ (red) consistently exhibits the highest relative proportion across all durations, while /ə/ and /u/ (orange and yellow) show slightly lower but similar trends. The two syllabic fricatives exhibit distinct patterns: /z/ (green) remains slightly lower than regular vowels but does not show a significant difference, whereas /ʒ/ (blue) consistently has the lowest relative proportion across all duration ranges, significantly lower than /a/, /ə/, and /u/.

4. Discussion

Our analysis of both spontaneous and read speech revealed robust differences in syllable nuclei duration, with syllabic fricatives consistently exhibiting shorter durations than regular vowels. Significant modulatory effects were observed: rising tones extended duration, falling tones reduced it, and certain consonantal contexts (e.g., /ts^h/) as well as sentence-initial and medial positions further compressed vowel length. Importantly, regular vowels appear more susceptible to these modulatory effects than syllabic fricatives, underscoring a differential impact of segmental and suprasegmental factors on vowel timing. The relative duration analysis highlights that the allocation of time within the syllable is systematically influenced by absolute duration. Very short vowels show a steep increase in their relative share, which then levels off as duration increases—indicating that while speakers maintain a relatively uniform temporal allocation for brief segments, differences among vowel types emerge with longer durations.

The results confirm our hypothesis: syllabic fricatives [ʒ, z] are acoustically shorter in Standard Mandarin compared to vocalic nuclei. While acoustic data cannot directly explain this durational specificity, we propose that aerodynamic maneuvers contribute to the shorter duration. This durational difference may reflect the pressure hierarchy tied to the production of voiced fricatives, which require precise control of airflow and vocal fold vibration. Future research should explore a wider range of syllable structures across Chinese dialects and languages, and incorporate aerodynamical measurements, such as airflow and pressure, to better understand non-articulatory constraints on temporal modulation in speech production.

5. Acknowledgements

This work was supported by the ANR project DIPVAR (ANR-21-CE38-0019), and also by the Laboratoire d'Excellence "Empirical Foundations of Linguistics" (LabEx EFL, ANR-10-LABX-0083). Jingyi Sun receives ongoing support from the China Scholarship Council (CSC Grant No.202208410095).

6. References

- [1] B. Karlgren, *Études sur la phonologie chinoise*. KW Appelberg, 1926, vol. 15.
- [2] S. Duanmu, *The phonology of standard Chinese*. OUP Oxford, 2007.
- [3] F. Dell, "Consonnes à prolongement syllabique en chine," *Cahiers de linguistique-Asie orientale*, vol. 23, no. 1, pp. 87–94, 1994.
- [4] W.-S. Lee and E. Zee, "Standard chinese (beijing)," *Journal of the International Phonetic Association*, vol. 33, no. 1, pp. 109–112, 2003.
- [5] S.-I. Lee-Kim, "Revisiting mandarin 'apical vowels': An articulatory and acoustic study," *Journal of the International Phonetic Association*, vol. 44, no. 3, pp. 261–282, 2014.
- [6] M. Faytak and S. Lin, "Articulatory variability and fricative noise in apical vowels." in *ICPhS*, 2015.
- [7] S. Foley, "The coarticulatory behavior of standard mandarin apical vowels," in *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS 2023)*, 2023.
- [8] S. Foley, B. Shao, and M. Faytak, "Relating frication to articulation in standard mandarin apical vowels," in *13th International Seminar on Speech Production*, 2024.
- [9] B. Shao and R. Ridouane, "On the nature of apical vowel in jixi-hui chinese: Acoustic and articulatory data," *Journal of the International Phonetic Association*, pp. 1–26, 2023.
- [10] J. J. Ohala, *The Origin of Sound Patterns in Vocal Tract Constraints*. New York, NY: Springer New York, 1983, pp. 189–216. [Online]. Available: https://doi.org/10.1007/978-1-4613-8202-7_9
- [11] —, "Aerodynamics of phonology," in *Proceedings of the Seoul International Conference on Linguistics*, vol. 92. Linguistic Society of Korea Seoul, 1997, p. 97.
- [12] P. Passy, *Étude sur les changements phonétiques et leurs caractères généraux*. Paris, France: Librairie Firmin-Didot, 1890.
- [13] N. Chomsky and M. Halle, *The Sound Pattern of English*, ser. Studies in English. Harper & Row, 1968.
- [14] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldia," in *Interspeech*, vol. 2017, 2017, pp. 498–502.
- [15] J. Sun, Y. Wu, N. Audibert, and M. Adda-Decker, "Création d'un corpus parallèle de styles de parole en mandarin via l'auto-transcription et l'alignement forcé," in *Actes des 35èmes Journées d'Études sur la Parole*, 2024, pp. 291–300.
- [16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [17] R. Luo, J. Xu, Y. Zhang, Z. Zhang, X. Ren, and X. Sun, "Pkuseg: A toolkit for multi-domain chinese word segmentation," *arXiv preprint arXiv:1906.11455*, 2019.
- [18] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glott International*, vol. 5, no. 9/10, pp. 341–347, 2001.
- [19] D. M. Bates, "lme4: Mixed-effects modeling with r," 2010.
- [20] M. E. Brooks, K. Kristensen, K. J. Van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Machler, and B. M. Bolker, "glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling," *The R journal*, vol. 9, no. 2, pp. 378–400, 2017.
- [21] B. Bolker, "Getting started with the glmmTMB package," *Cran. R-project vignette*, vol. 9, 2019.
- [22] J. Yuan, "The effects of speaking rate and intonation on the duration of tones in mandarin chinese," in *Speech Prosody 2012*, 2012.
- [23] J. Yang, Y. Zhang, A. Li, and L. Xu, "On the duration of mandarin tones," in *Interspeech*, 2017, pp. 1407–1411.
- [24] Y. Xu, "Contextual tonal variations in mandarin," *Journal of phonetics*, vol. 25, no. 1, pp. 61–83, 1997.