Speech Communication 146 (2023) 45–52

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Mandarin lexical tone duration: Impact of speech style, word length, syllable position and prosodic position

Yaru Wu[a,b,c,*], Martine Adda-Decker[b], Lori Lamel[c]

[a] CRISCO/EA4255, Université de Caen Normandie, 14000 Caen, France
[b] Laboratoire de Phonétique et Phonologie (UMR7018, CNRS-Sorbonne Nouvelle), France
[c] LISN/CNRS, UMR 9015, Université Paris-Saclay, 91405 Orsay, France

ARTICLE INFO

Keywords:
Mandarin
Tone duration
Number of syllable(s) in word
Syllable position
Prosodic position
Speech style
Large corpora, Continuous speech

ABSTRACT

This study aims to increase our knowledge of Mandarin lexical tone duration in continuous Mandarin speech. Related variation factors such as the number of syllable(s) in word, the position of syllable in word, its prosodic position and speech style were also explored. Large corpora of casual and journalistic speech (total ∼1000 hours) were used. More than 90% of the words (tokens) used in spoken Mandarin are monosyllabic and disyllabic words. In casual speech, 67% of the wordtokens are monosyllabic and 30% of the word-tokens are disyllabic. In journalistic speech, however, disyllabic words (49%) are more frequently used than monosyllabic words (45%). Tone 4 is the most frequently used tone among the four lexical tones in both casual (34%) and journalistic (36%) speech. Tone 1, Tone 2 and Tone 3 have similar occurrence frequencies in causal speech. Tone 3 appears to be the least frequently used tone in journalistic speech. With regard to tone duration, the results show that Tone 2 tends to have the shortest duration in causal speech and Tone 3 appears to have the longest duration in journalistic speech. Nonetheless, the studied variation factors (number of syllable(s) in word, position of syllable in word and prosodic position) are all found to influence the duration of Mandarin lexical tones, for both causal speech and journalistic speech. Tone durations in monosyllabic words appear to be closer to those of word-final syllables than to other syllable positions in multi-syllabic words. In terms of prosodic position, tone duration tends to increase with higher prosodic level in both casual and journalistic speech. Regardless of tone nature and speech style, the longest tone duration is in phrase-final position, followed by word-final and then word-medial position. Regional variety for tone duration is explored using casual speech productions from speakers of five major cities of North and South-East China, namely Beijing, Shanghai, Wuxi, Suzhou and Nanjing.

1. Introduction

During the past decades, studies on the duration of Mandarin tones have been attracting interest and have initiated debates as to which lexical tone bears the longest duration (e.g. Ho, 1976; Tseng, 1990; Chang, 2010). However, due to limitations related to available and explorable data combined with the lack of technological support, studies on the duration of Mandarin tones are still mostly based on small, more or less controlled datasets. This study aims to exploit the link between speech technology and linguistic research in Mandarin by studying the durations of Mandarin lexical tones in large speech corpora using tools from automatic speech recognition.

Mandarin Chinese (known as Standard Chinese or *Putonghua*) is a tonal language with four lexical tones: high (Tone 1), rising (Tone 2), low-dipping (Tone 3) and falling (Tone 4). The meaning of words depends on the lexical tone. For instance, “ma” (/ma/) could mean “mother” in Tone 1, “linen” in Tone 2, “horse” in Tone 3 and “scold” in Tone 4. In Mandarin, a morpheme, represented by a character in writing, can be a word itself or a part of a word. If a word is expected to have 4 characters in writing (i.e. 4 syllables in the canonical pronunciation), we consider it a four-syllable/terasyllabic word. Whereas most words in Mandarin carry a tone per syllable, the “neutral tone” (Tone 0) occurs in unstressed syllables of some words. This paper studies the duration of the four lexical tones in Mandarin Chinese and investigates several variation factors that could influence the duration of tones.

Ho (1976) observes that Tone 3 has the longest and Tone 4 the shortest duration among the four lexical tones in a study on isolated spoken words. More recent studies on continuous speech show diverging

* Corresponding author at: CRISCO/EA4255, Université de Caen Normandie, 14000 Caen, France.
E-mail addresses: yaru.wu@unicaen.fr (Y. Wu), martine.adda-decker@sorbonne-nouvelle.fr (M. Adda-Decker), lori.lamel@lisn.fr (L. Lamel).

https://doi.org/10.1016/j.specom.2022.11.001
Received 28 February 2021; Received in revised form 2 October 2022; Accepted 14 November 2022

results on tone duration. Tseng (1990) shows that Tone 3 is the longest among the four lexical tones in continuous Mandarin, in line with results in Ho (1976). Chang (2010) observes that Tone 2 tends to be the longest in sentence-final position. Yang *et al.* (2017), however, found that the four lexical tones have similar durations in continuous speech.

This study aims to analyze lexical tone duration in Mandarin and to contribute to the debates on tone duration. Different from previous studies on Mandarin tone duration, which focused on isolated monosyllabic words, on read sentences or on continuous speech of relatively small datasets, this study investigates tone duration of standard Mandarin in two large speech corpora, comprised of different speech styles. Furthermore, this paper aims to provide a better understanding of variation of tone duration by investigating the number of syllable(s) in word, the position of the syllable in the word, the prosodic position.

The position of a syllable in a word was found to be an important variation factor for vowel duration in French (Adda-Decker *et al.*, 2013). It is interesting to examine tone duration in Mandarin and verify whether the syllable position would have an influence on a tonal language as well. Chen (2006) investigated the durational adjustment of mono-morphemic four-syllable words in Standard Chinese and found an edge effect with the last syllable lengthened the most using laboratory data recorded in a controlled speech setting. Lai *et al.* (2010) analyzed a journalistic corpus of 28 speakers and found that the number of syllable (s) in a word could potentially influence tone duration. In this study, we further investigate the influence of syllable position in the word using a large-scale Mandarin speech corpus, where significant occurrences of three-syllable and four-syllable words are available for analyses.

Different acoustical or articulatory changes are found to correlate with prosodic position, allowing segments to be classified into hierarchically layered categories (Ladd *et al.*, 1986; Cho, 2016). Pierrehumbert *et al.* (1992) found VOTs of aspirated stops longer in a higher prosodic domain than in a lower one. Fougeron & Keating (1996) found that the articulation of the consonants [t] [n] depend on their prosodic position. An influence of prosodic position is found on segment duration (see Wightman *et al.*, 1992, Yoon *et al.*, 2007 for English and Fougeron & Jun (1998) for French). Yang & Wang (2002) found similar results in Mandarin using a 500-sentence corpus read by a 24-year-old female professional radiobroadcaster: pre-boundary syllable duration increases as the prosodic boundary degree increases. In this study, we intend to examine whether the findings of Yang & Wang (2002) on one speaker can be generalized. We study the influence of prosodic position on tone duration in Mandarin, with a focus on tone duration in word-internal, word-final and phrase-final positions. We expect to observe longer tone durations at higher prosodic levels.

To the best of our knowledge, the influence of speech style on tone duration has never been investigated using fluent speech. In the present paper, we investigate the influence of these factors in standard Mandarin using large corpora and verify whether the influence of the factors persists across the different speech styles.

Besides Standard Mandarin, the official language is often spoken with regional variations. As far as we know, few studies have covered this aspect of Mandarin Chinese. In this study, we also investigate regional varieties of Mandarin tone duration by speakers from 5 different cities, using a selected subset of the casual speech corpus. To date, only a few research studies were found on the regional variation of Standard Mandarin, let alone any large corpus-based study of the subject. Chang (2010), the study the most closely related to our analyses on regional variation, found that Tone 3 of Beijing Mandarin in isolation words was the only tone that was significantly different from the Taiwan Mandarin counterpart. We investigate whether tone 3 stays the tone that varies the most among regions in continuous speech and how those variations demonstrate among speakers of five different cities.

The methodology used in this study is detailed in Section 2 and the results on tone occurrences and durations are then presented in Section 3. A discussion of the results is developed in Section 4.

2. Methodology

In this section, we first introduce the corpora and the alignment method used for this study. Then, the description on duration measurements and investigated variation factors is provided. At the end of this section, we specify the statistical analyses applied for this study.

2.1. *Corpora and alignment*

Several corpora distributed by the Linguistic Data Consortium (Strassel *et al.*, 2006; Morris *et al.*, 2019) were used in this study, containing ~1000 hours of manually transcribed casual and journalistic speech (e.g., Huang *et al.*, 1998; Walker *et al.*, 2015; Meghan *et al.*, 2015). The journalistic corpus contains ~850 hours speech (8788281 word-tokens) containing mainly broadcast news and the causal speech corpus contains ~150 hours of speech (1710637 word-tokens) containing conversations between family members. The casual speech corpus contains mostly standard Mandarin speech but also a small subset of regional varieties of Mandarin speech. In our analyses on tone duration as a function of speech style, tone nature, number of syllable(s) in word, syllable position and prosodic position, only the standard Mandarin data is included.

As for our analyses on regional varieties of Mandarin tone duration, we used a small subset of the casual speech corpus, including conversational Mandarin produced by speakers from Beijing, Shanghai, Wuxi, Suzhou, Nanjing. We selected speakers from Beijing (given that it is comparable to standard Mandarin) and from 4 cities of south-east China. The 4 cities of south-east China were selected since they are the 4 most frequently observed cities in the portion of the database containing regional variety. To control for variability related to syllable structure and the nature of the adjacent consonant engaged in the rime of each syllable (i.e., consonant for which the duration is included as part of the tone duration; see more details of tone duration calculation in Section 2.2), our analyses on regional varieties focus on tone durations of _Vn and _Vŋ syllables.

All of the speech data was automatically segmented and annotated at the word and phone level using the LIMSI speech transcription system in forced alignment mode (Gauvain *et al.*, 2002; Adda-Decker *et al.*, 2000). The Mandarin pronunciation lexicon allows full access to phone level representations for each word carrying tones. Orthographic transcriptions guide the forced alignment with the system free to select the best matching pronunciation among the alternative pronunciations provided by the lexicon. Only a limited number of pronunciation variants were included in the baseline pronunciation lexicon and the average number of pronunciations per word is 1.05. The variants concern changes in phone (e.g. [b,p] for /b/) or in tone nature all 4 lexical tones are allowed for the character "啊", *particle showing affirmation*). The lexical tones are chosen during the forced alignment for words which allowed tone variants (less than 4%). The neutral tone was not integrated in the pronunciation lexicon and is therefore not modeled by the LIMSI speech transcription system. Pauses, hesitations, breath and noise are automatically detected. As described in Wottawa et al. (2018), the automatic alignment is realized using position-independent monophone acoustic models similar to those described in Chen et *al.* (2000) and Gauvain et *al.* (2002), as with such models the forced alignment has more freedom to select the best matching pronunciation variant. The acoustic model is a 3-state Hidden Markov acoustic model and the analyse has a 10 ms (frame) step resulting in a minimum phone duration of 30ms (Lamel *et al.*, 2011, Chen *et al.*, 2000). There are 72 phone models, 25 consonants and 11 vowels, each with 4 tones differentiated for each vowel. Three additional units are used to model silence, breath noises and filler words.

2.2. *Duration measurements*

Before presenting results on tone duration and factors of variations on tone duration, the relative frequencies of n-syllabic words and frequencies of each lexical tone were quantified as a function of speech style in our corpora.

Mandarin has relatively simple syllable structures. According to studies on generative phonology of Mandarin (Cheng *et al.*, 1973; Duanmu 1990), the longest Mandarin syllable contains four underlying elements CGVX (C = consonant; G = glide; V = vowel; X = offglide of a diphthong or a nasal consonant). In this study, we consider tone duration as the duration of the rhyme, starting from the onset of the vowel and through the end of the syllable (Howie 1976), that is to say, the duration of V+X of the word.

In continuous speech, the median segment duration is around 70ms and segments longer than 200ms typically corresponds to pauses or hesitations. Tone durations under 400 ms (since a tone segment could potentially cover both the vowel and the nasal segments) are included in this study. Fewer than 1% of the tone segments were excluded with the 400 ms threshold. Tone duration analyses were limited to words with 1- to 4- syllables and were normalized to z-scores.

Tone segments of three prosodic positions were differentiated: word-medial, word-final and phrase-final. Phrase-final tones refer to tones in the last syllable of an utterance preceding a pause ($\geq$ 100 ms)[1]. Word-final tones represent tone segments of the last syllable of words[2]. Word-medial tones refer to tones in a word internal position (see also Table 1). The analyses on the effect of prosodic position on tone duration were limited to polysyllabic words (2-4 syllabic words) since both “word-internal” and “word-final” positions are not meaningful for monosyllabic words. The two speech styles considered in our duration analyses on standard Mandarin are casual telephone conversations and journalistic speech.

2.3. Statistical analyses

Linear models were carried out in R (R Development Core Team, 2019) for the statistical analyses of tone duration as a function of the following variation factors: tone nature, number of syllable(s) in word, position of the syllable in the word, prosodic position and speech style[3]. Three separate models were used to explore tone duration in Standard Mandarin. Trisyllabic and tetrasyllabic words were grouped together for the statistical analyses. In the first model, tone nature, number of syllable(s) in the word, position of syllable in the word and speech style on tone duration were included as independent factors. Interaction between tone nature and speech style was integrated in this model. The second model was constructed to investigate the interaction between speech style and number of syllable(s) in the word and that between speech style and position of syllable in the word. For this second model, the number of syllable(s) in word, position of syllable in word and speech style on tone duration were included as independent factors. Interaction between speech style and number of syllable(s) in the word and that between speech style and position of syllable in the word were included as well. The third model was constructed to examine the effect of prosodic position. The dataset is smaller than that used in the first and the second model, given that specific prosodic positions were selected. We included prosodic position (word medial, word final and phrase final) as the independent factor in question. Tone nature, number of syllable(s) in the word and speech style were included as control variables. Interaction between speech style and tone nature and that between speech style and prosodic position in the word were included as well. In our analysis on regional varieties, we investigated _Vn and _Vŋ contexts and speaker origins (independent factors) and included tone nature, number of syllable(s) in the word and position of syllable in the word as control variables. Interaction between context and speaker origins was integrated in this model as well. Speech style is not concerned here, given that there is only continuous speech in this selected subset of our data. Post-hoc tests were performed (using *emmeans*) to perform pairwise comparison on different levels within the investigated factors for each of the models. Table 1 provides details on the variation factors.

3. Analyses and results

The occurrence frequencies of words and tones allow us to have a better understanding of how each analyzed category is distributed in the analyzed corpora. In this section, we first present the occurrence frequencies of words according to number of syllable(s) in the word and lexical tones as a function of speech style for our 1000-hour Mandarin speech corpora (Section 3.1). In non-tonal languages, the more frequent words tend to have shorter durations (see Wright 1979 for English). We therefore wonder if this is the case for tone duration. The frequency analyses of the lexical tones also allow us to verify whether there is a link between tone frequency and tone duration.

The duration of tones is analyzed as a function of speech style, tone nature, number of syllable(s) in word, position of syllable in word and prosodic position. The last part of this section presents the study of the duration of tones produced by speakers from five different cities in China.

3.1. Word and lexical tone occurrence frequencies

The frequencies of words as a function of number of syllable(s) in the word and speech style are presented, followed by a comparison of the occurrences of lexical tones as a function of speech style.

3.1.1. Word frequencies

Fig. 1 gives the distribution of words as a function of number of syllable(s) in the word and speech style. For both casual conversational and journalistic speech, more than 90% of the word (tokens) are monosyllabic and disyllabic words (97% for causal speech and 94% for journalistic speech). These results confirm that the most frequently used Mandarin words are short, with 57% (913583 word-tokens) and 45% (3942780 word-tokens) for monosyllabic words in casual and journalistic speech respectively and 40% (636388 word-tokens) and 49% (4310358 word-tokens) for disyllabic words for casual and journalistic speech respectively. 3% (46870 word-tokens) and 5% (402005 word-tokens) of trisyllabic words are found in causal speech and in journalistic speech respectively. Less than 6% of word-tokens containing more than 2 syllables per word are observed for both speech styles, with 3% (52496 word-tokens) and 6% (535143 word-tokens) for casual speech and for journalistic speech respectively. For both causal speech and journalistic speech, words longer than 4 syllables represent under 1% of all occurrences. Long words such as “印度尼西亚” (*Indonesia*) may exist in Mandarin Chinese, however, they are observed to be rather marginal.

As can be seen on the left panel, more monosyllabic words than disyllabic words are observed in the casual speech and more disyllabic words than monosyllabic words are observed on the right panel for journalistic speech. Less than 2% (112497 word-tokens) of tetrasyllabic words are observed in journalistic speech and even fewer tetrasyllabic words (< 1%, 194 word-tokens) are observed in casual speech. These results also suggest differences in word choices in different speech styles. The less formal the speech style is, the more monosyllabic words (e.g. “我”*I*; “你”, *you*) and the less dissyllabic words occur in our speech corpora (χ^2(1, N= 9803109) = 65116, p < 0.001). Similar patterns can be found for words with schwa in French: more than 70% of word-tokens with schwa are monosyllabic words covering few word-types (Wu *et al.*, 2020a).

[1] This value was determined after tests with pause durations of 50 ms, 100ms and 200ms. The three thresholds gave the same results as far as the influence of prosodic position is concerned.

[2] Tones at phrase-final positions are not included in this category.

[3] Linear mixed models were carried out, including speaker identity and item/word-type as random effects, but they did not converge due to the size of the datasets.

Table 1
Examined factors for tone duration. # stands for word boundary; ## for phrase boundary. Relevant syllables for the analyses on prosodic position are circled.

Factors	Levels	Examples
Tone nature	Tone 1	妈 (/ma1/, mother)
	Tone 2	麻 (/ma2/, linen)
	Tone 3	马 (/ma3/, horse)
	Tone 4	骂 (/ma4/, scold)
Number of syllable(s) in word	1syllable	九(nine)
	2 syllables	国家 (country)
	3 syllables (> 2 syllables)	消费者 (consumer)
	4 syllables (> 2 syllables)	画蛇添足 (gild the lily)
Syllable position in word	Last syllable (S_n)	画蛇添足 消费者 国家 九
	Last syl. – 1 (S_{n-1})	画蛇添足 消费者 国家
	Last syl. – 2 (S_{n-2})	画蛇添足 消费者
	Last syl. – 3 (S_{n-3})	画蛇添足
Prosodic position	Word-medial	消费者 (consumer)
	Word-final	消费者 #
	Phrase-final	消费者 ##
Speech style	Casual speech Journalistic speech	
Regional varieties	Standard Mandarin: not concerned 5-city Mandarin: Beijing, Shanghai, Wuxi, Suzhou, Nanjing	

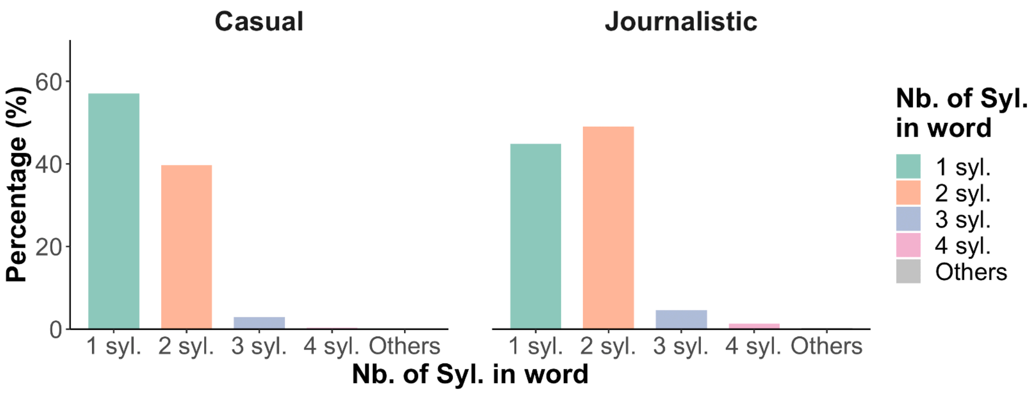

Fig. 1. Word-token frequencies as a function of the number of syllable(s) per word and speech style in standard Mandarin (1000h corpus: ~850h for journalistic speech and ~150h for casual speech).

In order to have a complementary view of the word-token frequencies reported in this paper, the percentages of word-types as a function of number of syllable(s) in the word are presented for each speech style in Fig. 2. These results show that the most frequently used word-types are disyllabic (68% in casual speech and 62% in journalistic speech). Note that we observe significantly fewer 3-syllable and 4-syllable words in casual speech than in journalistic speech (17% and 27% for casual and journalistic speech respectively; $\chi^2(1, N= 14786) = 16.706$, $p < 0.001$). This observation is consistent with findings in other languages (see Ryan & Sebastian (1980) and Yaguchi *et al.* (2010) for English; Wu

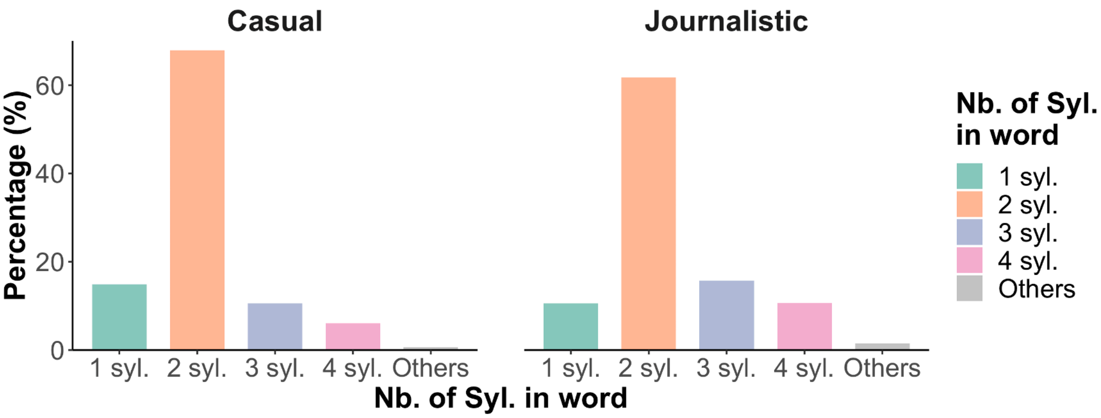

Fig. 2. Word-type frequencies as a function of the number of syllable(s) per word and speech style in standard Mandarin (1000h corpus: ~850h for journalistic speech and ~150h for casual speech).

et al. (2017), Adda-Decker & Lamel (2018) for French) and suggests a general tendency of employing longer words in a more formal speech setting in languages of different families.

3.1.2. Tone frequencies

Occurrence frequencies of the four lexical tones in standard Mandarin are presented in this section. Tone frequencies are shown as a function of tone nature and speech style.

Fig. 3 illustrates the distribution of each lexical tone in standard Mandarin for both casual and journalistic speech. It can be seen that the falling Tone 4 is the most frequently used tone, followed by the stable high Tone 1, for both casual speech (Tone 4: 34%, 980962 tone segments; Tone1: 28%, 795075 tone segments) and journalistic speech (Tone 4: 36%, 5130257 tone segments; Tone 1: 25%, 3561089 tone segments). Similar proportions of Tone 2 (19%, 546813 tone segments) and Tone 3 (19%, 561058 tone segments) are found in casual speech. As for the journalistic speech, Tone 2 comes the third (22%, 3168283 tone segments) and Tone 3 (17%, 2375490 tone segments) is the least frequent tone.

3.2. Duration of Mandarin lexical tones

The durations of lexical tones in standard Mandarin are analyzed in this section.

Normalized tone duration as a function of number of syllable(s) in the word and speech style is shown in Fig. 4. It can be seen that casual speech and journalistic speech show very different tone duration patterns. The figure suggests that Tone 1 has the longest duration among the four lexical tones. In this figure, Tone 3 is observed to be the longest among the four lexical tones in journalistic speech, regardless of number of syllable(s) in the word. However, this is not the case for casual speech. These findings suggest that the duration of tones depends on the speech style. Results of *emmeans* based on the relevant model suggest that all pairwise comparisons between tone nature and speech style are significant (p<0.001) except for “T2 casual - T4 journal” comparison.

The tone duration pattern for journalistic Mandarin is closer to previous findings on the duration of isolated tones than casual speech is. This might be due to the fact that speakers sometimes depend on written manuscript to speak and speakers in causal speech tend to shorten tone duration in production except for the stable high tone 1. The most frequently observed Tone 4 does not consistently have the shortest duration and the least frequent Tone 3 does not always have the longest duration. These results suggest that frequency alone is not sufficient to predict tone duration. These duration patterns do not take into account potentially important factors such as syllable position and prosodic structure, factors that are addressed in the next two sections. Linear regression results confirm the effect of tone nature on tone duration, while taking into consideration the effect of word length and syllable position.

3.3. Effect of the number of syllable(s) in a word and the syllable position on tone duration

In this section, the effect of number of syllable(s) in the word and that of the position of syllable in the word are analyzed for both casual and journalistic speech. We did not include 4-syllable words for causal speech in this analysis since less than 1% (194 word-tokens) of the words are tetrasyllabic as mentioned in Section 3.1.

Fig. 5 shows tone duration grouped by number of syllable(s) in the word, position of syllable in the word and the speech style. The four lexical tones are not shown individually here to save space, since analyses on each lexical tone shows the same pattern as is illustrated in Fig. 5. It can be seen that casual speech and journalistic speech show similar patterns with respect to the influence of number of syllable(s) in the word and the position of the syllable in the word. For both speech styles, monosyllabic words show tone durations closer to those of word-final syllables than to other syllable positions in polysyllabic words. Tone duration appears to be the longest in word-final syllables than in other syllable positions of polysyllabic words. Linear regression results confirm that syllable position affect tone duration for words of different length (here, 1-4 syllable words). Tone duration is significantly shorter in S_{n-2} [β = -0.061353; t = -4.092; SE = 0.014992] and in S_{n-1} [β = -0.045998; t = -3.129; SE = 0.014702] than in S_{n-3}. However, tones have significantly a longer duration in S_n than in S_{n-3} [β = 0.254854; t = 17.320; SE = 0.014714]. Post-hoc tests show that all pairwise comparisons among the four positions of syllable in word are significant (p < 0.001). An overall interaction between syllable position and speech style is not found in our data.

3.4. Effect of prosodic position on tone duration

The effect of prosodic position on tone duration is shown in Fig. 6 (causal speech on top panel; journalistic speech on the bottom). Tone duration appears to be longest at phrase-final position among the three prosodic positions, and this is the case for all four lexical tones and for both speech styles. Tone duration is the shortest in word-medial position, followed by word-final position. The figure suggests that tone duration increases with higher prosodic levels. The linear regression results confirm that tone durations are significantly longer in word-final position [β = 0.426214; t = 71.331; SE = 0.005975] and in phrase-final position [β = 1.352240; t = 190.025; SE = 0.007116] than that observed in word-medial position. Post-hoc tests show that all pairwise comparisons of prosodic levels are significant (p < 0.001). Results of *emmeans* based on the relevant model suggest that all pairwise comparisons between prosodic position and speech style are significant (p<0.001).

3.5. Regional varieties of tone duration

In this section, we investigate differences in tone duration according to speaker origin, namely Beijing, Shanghai, Wuxi, Suzhou and Nanjing.

Fig. 7 illustrates tone duration of _Vn and _Vŋ syllables grouped by

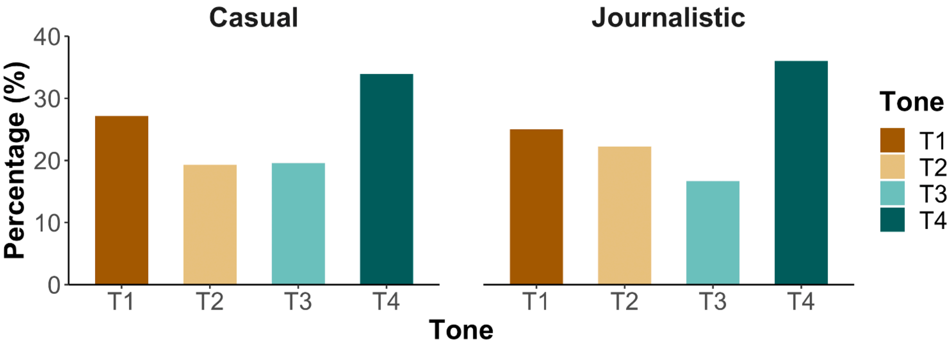

Fig. 3. Overall relative occurrence frequencies of Mandarin lexical tones for both speech styles.

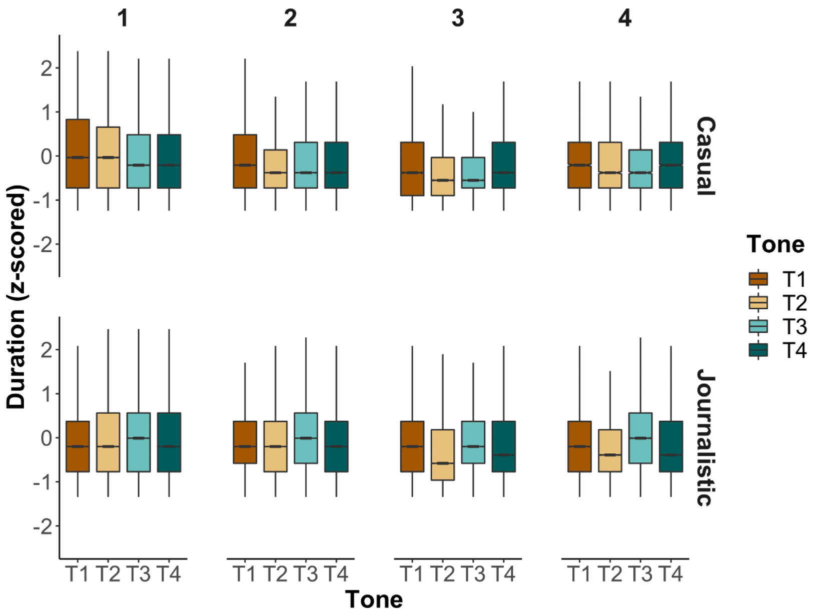

Fig. 4. Boxplot of the normalized tone durations in standard Mandarin as a function of the number of syllable(s) in the word (columns) for both casual and journalistic speech (rows).

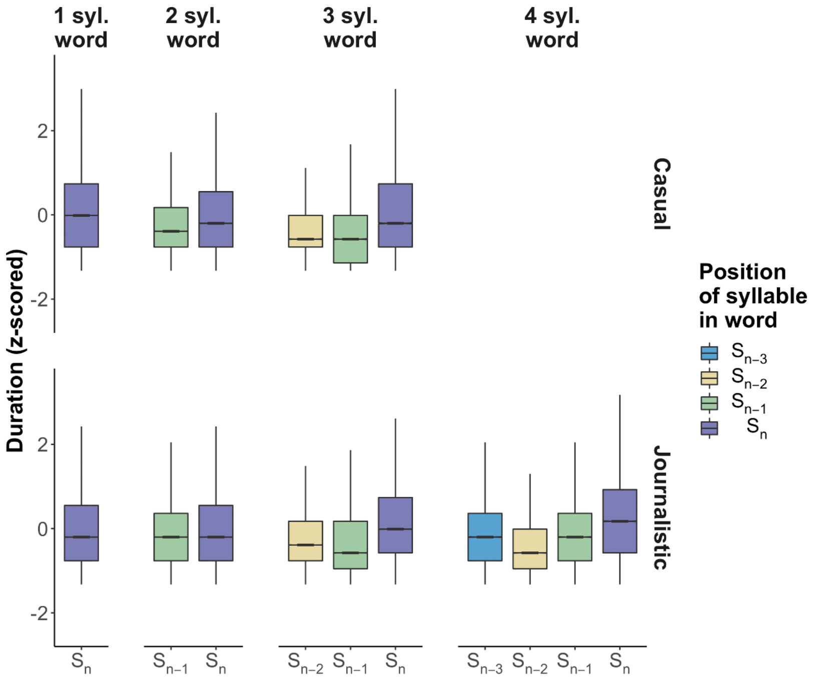

Fig. 5. Normalized tone durations (all tones pooled) as a function of number of syllable(s) in word, position of syllable in word and speech style.

tone nature for both speech styles. Tone duration tends to be longer for _Vŋ syllables than for _Vn syllables for all four lexical tones, which is a reassuring result given that the intrinsic duration of /ŋ/ is longer than that for /n/. Linear regression results confirm a significant difference between tone duration in _Vn and _Vŋ syllables, with shorter duration patterns observed for _Vn syllables [β = -0.15890; t = -14.285; SE = 0.01112].

Fig. 8 gives a more detailed illustration of the regional variation for tone duration of _Vn and _Vŋ syllables. These results show interesting patterns for _Vn and _Vŋ syllables. Tone duration tends to be more stable across tone natures for _Vŋ syllables than for _Vn syllables. It can be noted in the figure that, for Beijing, Shanghai and Nanjing, all tones tend to have similar duration for _Vŋ syllables (T1 ≈ T2 ≈ T3 ≈ T4) except for Wuxi (T3 > T1, T2, T4) and Suzhou (T1 > T2, T3, T4). Tone duration tends to be shorter for Nanjing [β = -0.10811; t = -3.243; SE = 0.03333] and Wuxi [β = -0.35951; t = -9.088; SE = 0.03956] speakers of Standard Mandarin than that observed for Beijing speakers. Interestingly, longer tone duration is observed for Shanghai speakers of Standard Mandarin than that observed for Beijing speakers [β = 0.07134; t = 2.803; SE = 0.02545]. No significant result is observed for Suzhou Mandarin. This suggests that Mandarin tone duration varies from region to region. This is different from what is observed in Chang (2010) on isolated tone production – here no specific pattern on the nature of the lexical tone (T1/T2/T3/T4) is observed for _Vn and _Vŋ syllables in continuous speech. No interaction between speaker origin (i.e. city) and tone nature is observed ($p > 0.05$).

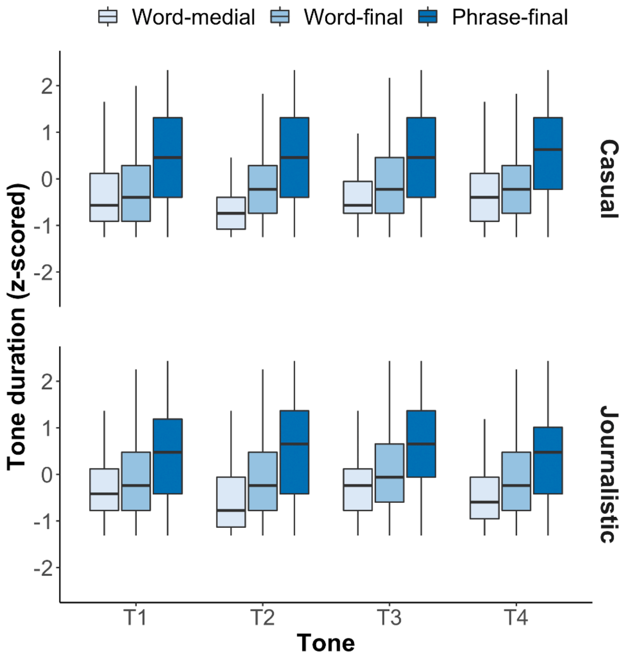

Fig. 6. Normalized tone duration as a function of tone nature and prosodic position: word-medial vs. word-final vs. phrase-final tone segments (top panel for casual speech; bottom panel for journalistic speech).

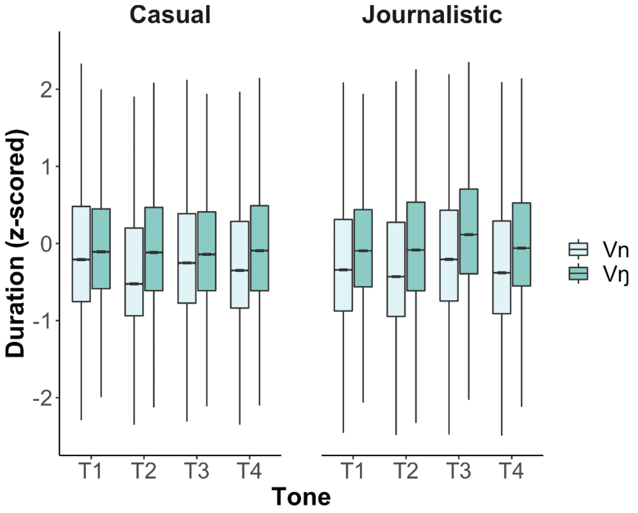

Fig. 7. Normalized tone duration on _Vn and _Vŋ syllables grouped by tone nature for each speech styles.

4. Discussion and Conclusions

This study investigated the duration of Mandarin lexical tones, here defined as the duration of the segment starting from the onset of the vowel and ending at the offset of the syllable. It extends the work presented at Interspeech 2020 (Wu *et al.*, 2020b) to the analysis of a much larger data set (1000 vs 220 hours), of a second speech style and also a first analysis of tone variation as a function of the speaker's dialect. The tone durations of the four lexical tones were examined as a function of tone nature, number of syllable(s) in the word, the position of syllable in the word, the prosodic position and the speech style. To our knowledge, this is the first time ever that linguistic study of Mandarin tones has been carried out on a 1000 hour corpus of continuous Mandarin speech.

We first quantified words according to their number of syllable(s) and speech style. More than 90% (casual speech 97%; journalistic speech 94%) of words used in spoken Mandarin are either mono- (casual speech 67%; journalistic speech 45%) or di-syllabic (casual speech 30%; journalistic speech 49%). Notably, more monosyllabic words than disyllabic words are found in casual speech; the inverse trend is found for the more formal, journalistic speech. Regarding the lexical tones, the falling Tone 4 is the most frequently observed tone in both causal (34%) and journalistic (36%) speech. Low-dipping Tone 3 (19%) and rising Tone 2 (19%) have similar frequencies in casual speech however in journalistic speech, rising Tone 2 (22%) is more frequent than low-dipping Tone 3 (17%), which is the least frequent tone in journalistic speech.

The four lexical tones in Mandarin do not always have similar durations and vary according to the number of syllable(s) in the word and the speech style, among other factors. Our results are different from those observed by Yang et al. (2017), who found all four lexical tones to have similar tone durations in <1 hour read speech and ~2 hours conversational speech. The difference in results might be due to difference in corpus size and the communication settings of the speech. In general, our results on tone duration in continuous speech suggest that it is not sufficient to discuss the duration of Mandarin lexical tones without taking into consideration different variation factors, such as those explored in this study. Different from studies of word production length and frequency for non-tonal languages, no direct correlation is found between tone frequency and tone duration in continuous Mandarin data.

Our results on the influence of number of syllable(s) in the word and the syllable position in the word suggest that tones in monosyllabic words tend to have durations close to that of word-final syllables of polysyllabic words for both speech styles. Independent of the number of syllable(s) in word, tone durations in word-final syllables tend to be the longest among all syllable positions in word, regardless of speech style. This finding is especially interesting because it is in line with vowel duration patterns as a function of number of syllable(s) in the word and position of syllable in French (Adda-Decker *et al.*, 2013), a language that is typologically very different from Mandarin. More importantly, this finding is in line with the edge effect found by Chen (2006) in mono-morphemic four-syllable words in Laboratory data. This match in results on experimental data and continuous speech data show that syllable position impact stone duration, regardless of the naturalness of the production setting.

The influence of prosodic structure on tone duration were analyzed: word-medial, word-final and phrase final positions. The results show that tone duration increases with prosodic level, regardless of tone nature and speech style. This finding is in line with other findings on the influence of prosodic position (Pierrehumbert *et al.*, 1992; Fougeron et al., 1995; Wightman *et al.*, 1992). Our results on tone duration are also consistent with those reported by Yang & Wang (2002) on the impact of a prosodic boundary on syllable duration in Mandarin.

Our results on _Vn and _Vŋ syllables suggest that regional variety influences tone duration, and that tone duration varies more in some varieties than others (e.g. patterns observed for _Vn *vs* _Vŋ syllables). Tone is known to be produced differently in isolated production and in continuous speech (see Ho, 1976 *vs* Tseng, 1990). Unsurprisingly, the difference found by Chang (2010) on tone duration in isolated tone production in Beijing and Taiwan Mandarin varieties is different from what we observed in continuous speech for Beijing, Shanghai, Wuxi, Suzhou and Nanjing varieties. No particular tone duration pattern is found for _Vn *vs*_Vŋ in these five varieties. This might be due to the language distances between the official language (here, Mandarin) and the local language spoken in these areas.

In short, our results suggest that, in continuous speech, the duration of Mandarin lexical tones not only varies according to tone nature (Tone 1 *vs* Tone 2 *vs* Tone 3 *vs* Tone 4), but also varies according to other variation factors, namely speech style, number of syllable(s) in word, position of syllable in word, prosodic position and the origin of the speakers.

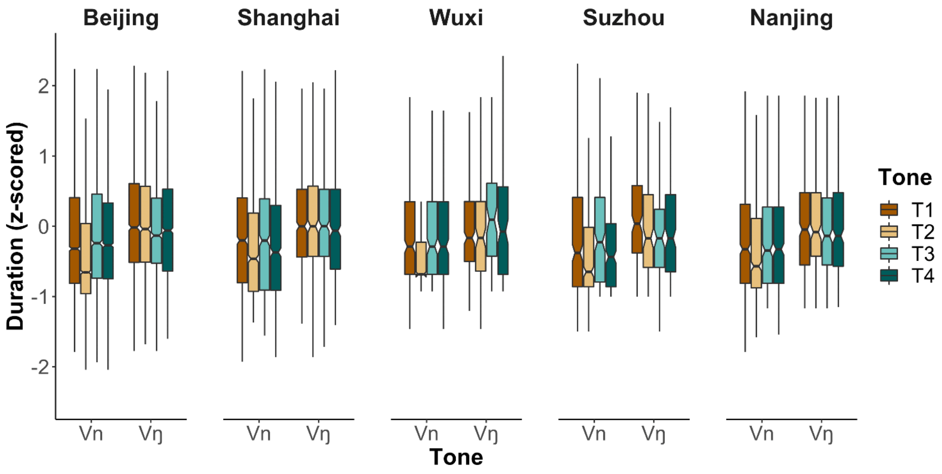

Fig. 8. Normalized tone duration on _Vn and _Vŋ syllables according to tone nature and _Vn and _Vŋ syllables.

Declaration of Competing Interest

Members of the following three laboratories: CRISCO/EA4255 - Université de Caen Normandie; LISN/CNRS - Université Paris-Saclay (former LIMSI/CNRS); Laboratoire de Phonétique et Phonologie (UMR7018, CNRS-Sorbonne Nouvelle), France

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083).

References

Adda-Decker, M., Lamel, L., 2000. The use of lexica in automatic speech recognition. Lexicon Development for Speech and Language Processing. Springer, Dordrecht, pp. 235–266.
Adda-Decker, M., Lamel, L., 2018. Discovering speech reductions across speaking styles and languages. Rethink. Reduct. 4, 101–128. De Gruyter Mouton.
Adda-Decker, M., Gendrot, C., Snoeren, N., & Nguyen, N. (2013). Apport du traitement automatique à l'étude des voyelles.
Chang, C.Y., 2010. Dialect differences in the production and perception of Mandarin Chinese tones. The Ohio State University (Doctoral dissertation.
Chen, L., Lamel, L., Adda, G., Gauvain, J.L., 2000. Broadcast news transcription in Mandarin. In: Sixth International Conference on Spoken Language Processing.
Chen, Y., 2006. Durational adjustment under corrective focus in Standard Chinese. J. Phon. 34 (2), 176–201.
Cheng, Chin-Chuan., 1973. A synchronic phonology of Mandarin Chinese. Monographs on Linguistic Analysis. The Hague: Mouton. *no. 4*.
Cho, T., 2016. Prosodic boundary strengthening in the phonetics–prosody interface. Lang. Linguist. Compass 10 (3), 120–141.
Development Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria [Online] Available. https://www.R-project.org/.
Duanmu, S., 1990. *A formal study of syllable, tone, stress and domain in Chinese languages* (Doctoral dissertation. Massachusetts Institute of Technology.
Fougeron, C., Jun, S.A., et al., 1998. Rate effects on French intonation: Prosodic organization and phonetic realization. J. Phon. 26 (1), 45–69.
Fougeron, C., Keating, P.A., 1995. Demarcating prosodic groups with articulation. J. Acoust. Soc. Am. 97 (5), 3384. -3384.
Fougeron, C., Keating, P.A., 1996. Articulatory strengthening in prosodic domain-initial position. UCLA Working Papers in Phonetics, 61–87.
Gauvain, J.L., Lamel, L., Adda, G., 2002. The LIMSI broadcast news transcription system. Speech Commun. 37 (1-2), 89–108.
Ho, A.T., 1976. The acoustic variation of Mandarin tones. Phonetica 33 (5), 353–367.
Howie, J.M., Howie, J.M., 1976. Vol. 18. Acoustical studies of Mandarin vowels and tones. Cambridge University Press.
Huang, S., Liu, J., Wu, X., Wu, L., Yan, Y., Qin, Z., 1998. Mandarin broadcast news speech (hub4-ne). Linguist. Data Consort.
Ladd, D.R., 1986. Intonational phrasing: the case for recursive prosodic position. Phonology 3, 311–340.
Lai, C., Sui, Y., Yuan, J., 2010. A Corpus Study of the Prosody of Polysyllabic Words in Mandarin Chinese. In: Speech Prosody 2010-Fifth International Conference.
Lamel, L., Gauvain, J.L., Le, V.B., Oparin, I., Meng, S., 2011. Improved models for Mandarin speech-to-text transcription. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4660–4663.
Morris, A., Antonishek, B., Li, X., Strassel, S., 2019. *HAVIC MED Progress Test–Videos, Metadata and Annotation*. Linguistic Data Consortium. University of Pennsylvania.
Pierrehumbert, J., Talkin, D, 1992. Lenition of/h/and glottal stop. In: Docherty, G., Ladd, DR (Eds.), Papers in Laboratory Phonology II, pp. 90–117 by.
Ryan, E.B., Sebastian, R.J., 1980. The effects of speech style and social class background on social judgements of speakers. Br. J. Soc. Clin. Psychol. 19 (3), 229–233.
Strassel, S.M., Cieri, C., Cole, A., DiPersio, D., Liberman, M., Ma, X., Maeda, K., 2006. Integrated Linguistic Resources for Language Exploitation Technologies. LREC 185–190.
Tseng, C.Y., 1990. An acoustic phonetic study on tones in Mandarin Chinese. In: Chung yang yen chiu yüan li shih yü yen yen chiu so, 94.
Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M., Price, P.J., 1992. Segmental durations in the vicinity of prosodic phrase boundaries. J. Acoust. Soc. Am. 91 (3), 1707–1717.
Wottawa, J., Amazouz, D., Adda-Decker, M., Lamel, L., 2018. Studying vowel variation in French-Algerian Arabic code-switched speech. In: Interspeech 2018. ISCA, pp. 2753–2757.
Wright, C.E., 1979. Duration differences between rare and common words and their implications for the interpretation of word frequency effects. Mem. Cognit. 7 (6), 411–419.
Wu, Y., Adda-Decker, M., Fougeron, C., Lamel, L., 2017. Schwa realization in French: using automatic speech processing to study phonological and socio-linguistic factors in large Corpora. In: Proc. Interspeech 2017, pp. 3782–3786.
Wu, Y., Adda-Decker, M., Lamel, L., 2020a. Schwa deletion in word-initial syllables of polysyllabic words: investigations using large French speech Corpora. J. Monoling. Biling. Speech 2 (2), 269–289.
Wu, Y., Adda-Decker, M., Lamel, L., 2020b. Mandarin lexical tones: a corpus-based study of word length, syllable position and prosodic position on duration. Proc. Interspeech 2020, 1908–1912.
Yaguchi, M., Iyeiri, Y., Baba, Y., 2010. Speech style and gender distinctions in the use of very and real/really: An analysis of the Corpus of Spoken Professional American English. J. Pragmat. 42 (3), 585–597.
Yang, Y., Wang, B., 2002. Acoustic correlates of hierarchical prosodic boundary in Mandarin. In: Speech Prosody 2002, International Conference.
Yang, J., Zhang, Y., Li, A., & Xu, L. (2017). On the duration of mandarin tones. In *INTERSPEECH* (pp. 1407-1411).
Yoon, T.J., Cole, J., Hasegawa-Johnson, M., 2007. On the edge: acoustic cues to layered prosodic domains. Proceed. ICPhS 16, 1264–1267.