Timing as an information carrier in speech

Chengxia Wang

A dissertation submitted to University College London

for the degree of

Doctor of Philosophy

Department of Speech, Hearing and Phonetic Sciences

Division of Psychology and Language Sciences

University College London (UCL)

May 2024

Dedication

I would like to dedicate my

thesis to my family and friends.

Declaration

I, Chengxia Wang, confirm that the work presented in this dissertation is my own. Where information has been derived from other sources, I confirm that this has been indicated in the dissertation.

Chengxia Wang

Abstract

In this thesis I investigated a number of timing related issues in speech from an information-theoretical perspective, based on the view that they arise either from communicative functions or articulatory mechanisms rather than for timing's own sake. I first examined the widely claimed hypothesis that languages of the world are either stress-timed, syllable-timed or mora-timed, by checking for evidence of at least tendencies toward isochrony in English and Mandarin, two languages alleged to be stress-timed and syllable-timed, respectively. The results show an absence of any tendency toward isochrony of stress groups in English, but a small tendency toward both isochrony of syllables and isochrony of phrases in Mandarin. Thus, the rhythm class hypothesis is argued to be not just weak, but untenable. In addition, I showed that the isochrony difference between the two languages is likely because the high functional load of English segments prevents the use of phrase-internal duration for phrase marking in addition to final lengthening. My further examination of boundary marking in the two languages also revealed a major difference in how prosodic boundaries above the phrase level are marked: Pre-boundary lengthening stops increasing beyond break index 2 in Mandarin, yet the increase is continuous in English. But this cross-language difference is evened out when final lengthening and silent pause are combined into an index of cross-boundary temporal distance, as Mandarin was found to mainly use silent pause to mark higher boundaries beyond B2. Finally, the information-theoretical perspective also sheds light on the so-called Part of Speech (POS) effect,

according to which nouns are known to be longer than verbs and function words. Results show that it is word frequency that has the most direct effect on duration, while effects of POS are likely a by-product of word frequency. The is consistent with information theory. That is, speakers may be under pressure to convey as much information as possible in a given amount of time, and this pressure would lead to each word being assigned as little time as possible. High frequency words can afford to have less time and thus less full articulation due to their predictability.

Impact statement

This thesis has the potential to improve our understanding of speech timing. By investigating three timing related issues: (1) if there is a tendency towards isochrony; (2) whether temporal distance can be used to distinguish boundaries in different languages; and (3) if there is a Part of Speech effect on duration, this dissertation presents a more complete picture of speech timing.

Although it has been known for a long time that there is no clear evidence of isochrony in any language and researchers are no longer looking for direct evidence of it, the notion that languages are divided into rhythm classes based on timing remains widespread, and it continues to drive rhythm-related research. Chapter 2 presents evidence that there is no tendency toward isochrony in English but Mandarin shows a weak tendency toward isochrony. The research method of comparing segment and syllable compressibility in English and Mandarin can be used to investigate rhythm in other languages pairs.

The key finding in Chapter 3 is that in English, the duration of pre-boundary syllables in English increases linearly with break index, whereas in Mandarin, the duration increase ends after break index 2. Since this demonstrates a significant difference of these two languages on boundary marking, the results can be directly applied to second language learning.

By comparing noun and verb homophones in designed sentences, Chapter 4 can facilitate our understanding of the reasoning for using POS as an essential input feature in the training process in both speech recognition and speech synthesis. The results show that there is no significant POS effect on duration, but there is a significant frequency effect on duration. Considering POS does show a difference in mean duration between nouns and verbs in my data, in the same direction as the frequency effect, it is possible to draw the conclusion that this weak POS effect is derived from a more primitive effect of frequency. These results can be used to improve performance in modelling.

Acknowledgements

Let me begin by expressing my gratitude to my principal supervisor, Yi Xu, for guiding me especially when I ran into challenges in my research. I benefited a lot from our weekly meetings. He has a lot of amazing research ideas. On a personal level, he offered me a lot of encouragement and advice, especially during the pandemic, which was a difficult time for many people.

Thanks to my subsidiary supervisor, Mark Huckvale, for providing useful insights for my research during upgrade. I would like to thank my examiners, Laurence White and Bronwen Evans, for their insightful comments. Jinsong Zhang (Beijing) inspired me to pursue a PhD in Speech Sciences. I will never forget our first meeting. I would especially thank Hao Liu for his help with my application to the PhD programme at UCL.

Sincere thanks to my friends and colleagues at UCL, including Anqi, Clara, Cris, Florian, Giulia, Gwijda, Howon, Julie, Nadine, Qing, Rachel, Shego and Yue.

My church friends have made my life in London much easier: Angela, Caleb, Colleen, Graham, Hannah, Matilda, Pauline, Sara and Seth. It was an honour to be a member of the church choir and I enjoyed many of the church events.

Many thanks to my friends Ashish, July and Leyuan for their help on math and programming; and to Jenn-Yeu Chen for offering me a two-month research assistant position at National Taiwan Normal University.

I want to thank my mother and brother for always believing in me and encouraging me to pursue my dreams, my best friend Qingcai for always being there for me, as well as my boyfriend Tristan for all his loving support.

Finally, I would like to express my gratitude to the China Scholarship Council and University College London for their financial support.

UCL Research Paper Declaration Form

referencing the doctoral candidate’s own published work(s)

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- *have been uploaded to a preprint server*
- *are in submission to a peer-reviewed publication*
- *have been published in a peer-reviewed publication, e.g. journal, textbook.*

This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)

a) **What is the title of the manuscript?**

Functional timing or rhythmical timing, or both? A corpus study of English and Mandarin duration

b) **Please include a link to or doi for the work**

https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.869049/full

c) **Where was the work published?**

Frontiers in Psychology

d) **Who published the work?** (e.g. OUP)

Frontiers Media SA

e) **When was the work published?**

2023

f) **List the manuscript’s authors in the order they appear on the publication**

Chengxia Wang, Yi Xu, Jinsong Zhang

g) **Was the work peer reviewed?**

Yes

h) **Have you retained the copyright?**

Yes

i) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If ‘Yes’, please give a link or doi)

No

If ‘No’, please seek permission from the relevant publisher and check the box next to the below statement:

☒

*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)

   a) **What is the current title of the manuscript?**

   Click or tap here to enter text.

   b) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if ‘Yes’, please give a link or doi)

   Click or tap here to enter text.

   c) **Where is the work intended to be published?** (e.g. journal names)

   Click or tap here to enter text.

   d) **List the manuscript’s authors in the intended authorship order**

   Click or tap here to enter text.

   e) **Stage of publication** (e.g. in submission)

   Click or tap here to enter text.

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

   Chengxia Wang and Yi Xu developed the theory. Chengxia Wang performed the analysis. Yi Xu supervised the project. Yi Xu provided BRC corpus and Jinsong Zhang provided ASCCD corpus. Chengxia Wang and Yi Xu wrote the manuscript with input from all authors. All authors contributed to the article and approved the submitted version.

4. **In which chapter(s) of your thesis can this material be found?**

   Chapter 2, Chapter 3

5. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

Candidate

Chengxia Wang

Date:

08/05/2024

Supervisor/ Senior Author (where appropriate)

Yi Xu

Date:

08/05/2024

List of publications

Wang, C., Xu, Y., & Zhang, J. (2023). Functional timing or rhythmical timing, or both? A corpus study of English and Mandarin duration. *Frontiers in Psychology*, *13*, 869049.

Wang, C., Xu, Y., & Zhang, J. (2022). The invalidity of rhythm class hypothesis. In *Proceedings of the International Conference on Speech Prosody* (Vol. 2022, pp. 347-351). International Speech Communication Association (ISCA).

Wang, C., Xu, Y., & Zhang, J. (2019). Mandarin and English use different temporal means to mark major prosodic boundaries. *Corpus*, *2*, 1.

Wang, C., Zhang, J., & Xu, Y. (2018, June). Compressibility of segment duration in English and Chinese. In *Proceedings of Speech Prosody 2018* (Vol. 9, pp. 651-655). International Speech Communication Association (ISCA).

Wang, C., & Xu, Y. (2017). Effects of part of speech: Primitive or derived from word frequency?. *ExLing2017*.

Table of contents

List of Figures

List of tables

1 Introduction

While it may seem obvious that human language is a communication system, it is far from clear how closely various aspects of speech are related to this fact. Even whether language is well designed for communication has been questioned. For example, Chomsky (2002) has argued that the use of language for communication is a kind of epiphenomenon: “If you want to make sure that we never misunderstand one another, for that purpose language is not well designed, because you have such properties as ambiguity. If we want to have the property that the things that we usually would like to say come out short and simple, well, it probably doesn’t have that property. Chomsky (2002, p. 107)”

One popular theory that does seem to seriously consider the importance of communication is the theory of hyper- and hypo-articulation (H&H) by Lindblom (1990). The theory is based on the principle of economy of effort (or ease of articulation), according to which speech is maximized for the economical use of articulatory effort. It states that speakers vary their output along a continuum of hyper- and hypo-speech based on the principle that “unconstrained, a motor system tends to default to a low-cost form of behaviour” (Lindblom, 1990, p.413). It therefore views speech articulation as a highly energy costly motor activity. Lindblom used energy saving to explain the phenomenon of undershoot (Lindblom, 1963), which is when articulators are unable to move in place, resulting in undershooting the target, if there is insufficient time. This is regarded as hypo-speech. While when a word conveys significant information and requires differentiation, the speaker will put more effort into its articulation, spending more time on its phonetic production; this is considered hyper-speech.

Some scholars have proposed using time pressure as an alternative explanation for undershoot to energy saving. Tiffany (1980, p. 907) came to the conclusion that “in some senses we normally speak about ‘as fast as we possibly can,’ at least in the production of full canonical utterances.” Studies of F0 production show that the maximum speed of pitch change is often approached (Xu and Sun, 2002; Kuo et al., 2007; Xu and Wang, 2009). Xu and Prom-on (2019) argued that Lindblom’s H&H theory is wrong, as phonetic undershoots are frequently due to the limit of articulatory speed rather than economy of effort. Xu and Prom-on (2019) then proposed maximum rate of information, implying that in speech production there is a pressure to transmit as much information as possible in a given amount of time. This new principle considers duration as a critical encoding dimension, and lengthening and shortening are important cues to mark certain communicative functions.

1.1 Information theory

The difficulty of using communication as an explanation factor could be resolved, however, by considering Shannon’s (1948) “Information Theory”, whose application has led to the invention of the internet, computer networks, CDs, data storage, wireless systems, mobile phones, MP3s, JPEGs, etc. Figure 1 displays a schematic diagram of a general communication system. Here the encoding system of message is of critical importance. If a piece of message needs to be transmitted, the first step is to encode it.

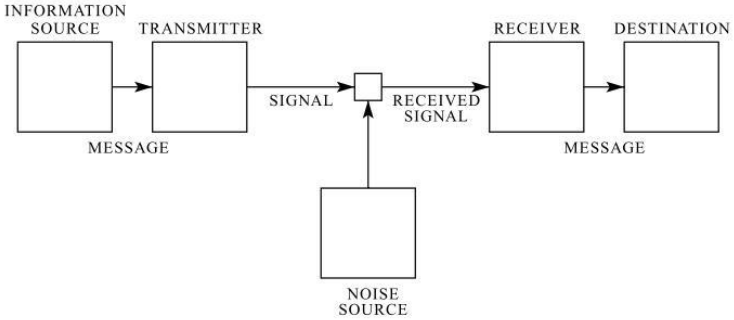


Figure 1. A schematic diagram of a general communication system (Shannon ,1948). First, a source of information generates a message for transmission. Next, the message is turned into a signal, which is then sent to a receiver through a channel. The signal is then decoded by the listener to get the message.

Frequency is a main point in reducing required channel capacity when encoding information in information theory (Shannon, 1948). Short codes are used to represent messages with high probability, while long codes are used to indicate messages with low probability. For instance, the letter E in English appears more frequently than Q, the sequence TH more frequently than XP, etc. In telegraph, the shortest channel symbol, a dot, is thus used to represent the most common English letter E, whereas the less common letters, Q, X, Z are represented by longer sequences of dots and dashes. Due to the availability of this structure, one can make a saving in time (hence channel capacity) by appropriately encoding the message sequences into signal sequences (Shannon, 1948).

Language exhibits similar trends. According to Zipf's law, word length has an inverse relationship with frequency (Zipf, 1935). He believes that this is a result of communicative efficiency. If we consider the alphabet as a language code, we would use longer sequences to code less frequent meanings and shorter sequences to code frequent meanings. If this is the case, frequency should be able to predict word length in some experimental data.

Among the major issues addressed by the theory is that of capacity of information transmission, i.e., how much information can be transmitted through a given channel within a given amount of time, or the rate of information transmission. Shannon's theory has proposed a mathematical solution to the problem of how the amount of information can be represented and calculated, as well as how much information can be transmitted through any given channel. According to this theory, the capacity of a channel is defined by the rate of transmission, which means that time is a core dimension of communication.

1.2 Articulatory-functional perspective of timing

Timing is an integral Part of Speech. Our understanding of timing is still limited although there is a lot of research on timing-related topics. Xu (2009) proposed an articulatory-functional perspective of timing, according to which timing in speech is either obligatory—timing as obligated by articulation, or informational—timing that conveys communicative meanings.

The human articulatory apparatus is constrained by physical laws. It takes time to manipulate the state of the articulatory system during speech, and a portion of that time is directly due to the minimum duration of articulatory movements (Xu, 2009). This is consistent with intrinsic duration of segments (Klatt, 1976; Lehiste, 1972; Port, 1981) which refers to the fact that each vowel or consonant has its own relative duration independent of other factors. Because this type of speech timing is obligatory, it cannot be used for information coding. From the articulatory-functional perspective, timing can be actively controlled. White (2014) also suggested a functional approach to prosodic speech timing, aiming to identify the effects that serve a linguistic communicative purpose.

There are several well-known functions that can influence timing, such as stress, focus and boundary. Timing is an important cue for word stress. In English, Fry (1955, 1958) carried out a series of experiments to investigate the impact of certain physical cues on the perception of stress in English. He showed that while pitch is by far the most important factor, duration is more important than intensity. The stressed/unstressed duration ratio is 2.18:1 according to Crystal and House (1988).

Timing is also important in marking focus. Focus is a function that serves to highlight a certain piece of information in comparison to information already shared by the conversation participants (Bolinger, 1972; Gussenhoven, 2008; Ladd, 1996; van Heuven, 1994). Syllable duration increases significantly under focus (Turk & Shattuck-Hufnagel, 2000; Xu, 1999; Xu & Xu, 2005).

Boundary is a function that allows continuous speech to be broken down into smaller chunks for ease of auditory comprehension (Lehiste, 1972; Cutler et al., 1997; Schafer et al., 2000; Xu, 2019). It has been discovered that boundary is associated with pre-boundary lengthening, which refers to the phenomenon that syllables and their component segments before a prosodic boundary are longer than they would be in other contexts (Lehiste, 1972; Nakatani et al., 1981; Shattuck-Hufnagel & Turk, 1996; Xu & Wang, 2009). It has also been reported that the duration of consonants is longer in word-initial position than that in word-medial position (Oller, 1973; Cooper, 1991; Fougeron and Keating, 1997; White and Turk, 2010). Fougeron and Keating (1997) showed that initial consonant /n/ in English follows the prosodic hierarchy, with greater initial lengthening after larger boundaries. There is increasing evidence that duration provides the strongest and most reliable cue for boundary (Wagner, 2005; Xu, 2009; Xu, 2019). Pitch reset (Ladd, 1988; Swerts, 1997) is also evidence of boundary, but it only occurs in limited situations and is often a result of other communicative functions.

It seems clear that the durational contrast for word stress, focus and boundary is communicative, but the communicative functions of some other timing patterns are unclear. Timing in speech rhythm, timing to mark major prosodic boundaries across languages and Part of Speech effect on duration are some of the examples that will be discussed in this thesis. In the following sections, these topics will be reviewed in greater detail.

1.3 Timing in speech rhythm

> "… every language in the world is spoken with one kind of rhythm or with the other. In the one kind, known as a syllable-timed rhythm, the periodic recurrence of movement is supplied by the syllable-producing process: the chest-pulses, and hence the syllables, recur at equal intervals of time—they are isochronous. … In the other kind, known as a stress-timed rhythm, the periodic recurrence of movement is supplied by the stress-producing process: the stress-pulses, and hence the stressed syllables, are isochronous."
>
> Abercrombie, 1967

Ever since the classic works of Pike (1945) and Abercrombie (1964a, 1964b, 1967), one approach to categorising languages according to their speech rhythm has been to describe them as either stress-timed, syllable-timed or mora-timed (Arvaniti & Rodriquez, 2013; Ramus et al., 1999). As illustrated in Figure 2, in a stress-timed language, inter-stress intervals are constant, hence, isochronous, whereas in a syllable-timed or mora-timed language, successive syllables or morae are equal in duration (Abercrombie, 1964a, 1967; Pike, 1945). Languages like English, Russian, Arabic, in fact, most Germanic and Slavonic languages, are deemed stress-timed, while French, Telugu, Yoruba, and most Romance languages are believed to be syllable timed (Ladefoged, 1975; Pike, 1945; Rubach & Booij, 1985), and languages like Japanese and Tamil are regarded as mora-timed (Bertinetto, 1989; Port et al., 1987; Steever, 1987).

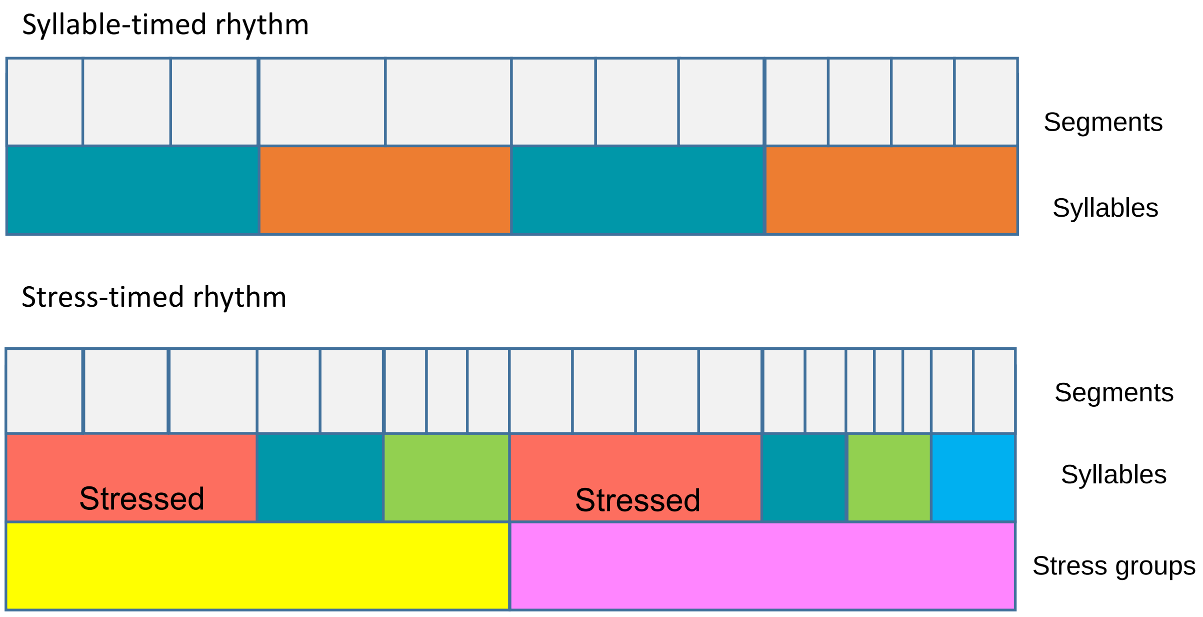


Figure 2. Isochrony in syllable-timed and stress-timed languages.

Experimental investigations, however, have been unable to find evidence of isochrony in either stressed-timed, syllable-timed or mora-timed languages. For stress timing, time spans between primary stresses in English did not cluster around some average value (Shen & Peterson, 1962). In “The North Wind and the Sun” read by David Abercrombie, inter-stress intervals showed no marked regularity (Uldall, 1971). In the six languages examined by Roach (1982), stress-timed ones exhibited a wide range of percentage deviations in inter-stress intervals. In fact, a proportional relationship is found between the number of syllables and duration of inter-stress intervals. As the number of segments increases, foot duration showed a clear tendency to increase (O’Connor, 1968), and the relationship between the number of intervening unstressed syllables and the inter-stress interval for real words in sentence context is linear (Lea, 1974). For syllable timing, in French a twelve-syllable sequence is not twice as long as a six-syllable sequence (Wenk & Wioland, 1981). In Spanish, syllable duration varied with the complexity of syllable structure, stress and position (Borzone de Manrique & Signorini, 1983). Pointon (1980), after reviewing the findings of a number of studies, concluded that “Spanish has no regular rhythm in the sense of an isochronous sequence of similar events, be they syllables or stress.” For mora timing, Warner and Arai (2001) found no evidence that speakers adjust the duration of segments within a mora in spontaneous speech in Japanese that would make morae equal in duration.

It is clear that the surface isochrony at syllable, stress group or mora level is not observed. An interesting question is, what if there is still a tendency toward isochrony, that is, speech rhythm is a top-down control of timing aiming at isochrony? To answer this question, it is important to consider that duration is also influenced by many other factors, including the position of the interval within utterance, specific types of syllables they contain, manner of articulation of consonants, and complexity of syllable structure, stress and position. Hence, assessing the presence of a tendency toward isochrony involves examining whether there is any duration variation left to show any tendency toward constant duration of syllables, stress groups or moras when the effects of all the other factors are ruled out.

To achieve isochrony, however, the components of the isochronous units need to be flexible in duration, as recognized by Pike (1945), Abercrombie (1967) and Roach (1982). In stress-timed language, “since the rhythm units have different numbers of syllables, but a similar time value, the syllables of the longer ones are crushed together, and pronounced very rapidly, in order to get them pronounced at all, with that time limitation” (Pike, 1945). As illustrated in Figure 2, for syllables to be isochronous, segmental duration needs to be flexible to compensate for changes in the number of segments in a syllable; and for stress groups to be isochronous, syllable duration needs to be flexible to accommodate the number of syllables in a stress group. Most importantly, without flexible segmental duration, stress timing is also impossible, because syllables have to vary in duration in the first place to guarantee equal inter-stress intervals (Hoequist Jr, 1983).

The problem of lack of evidence for isochrony seems to have been brushed away, however, by the proposal of the rhythm metrics since the late 1990s, starting from Ramus et al. (1999). These metrics are shown to be able to quantify the rhythm class of languages using consonantal and vocalic variability. The main measurements are %V (the proportion of vocalic intervals in an utterance), ∆V (the standard deviation of vocalic intervals within an utterance), ∆C (the standard deviation of consonantal intervals within an utterance) (Ramus et al., 1999), VarcoC (Standard deviation of consonantal intervals divided by mean and multiplies 100), VarcoV (Standard deviation of vocalic intervals divided by mean and multiplies 100) (Dellwo, 2006; Dellwo & Wagner, 2003), and the pairwise variability indices nPVI and rPVI (Pairwise Variability Index in their measurements on successive vocalic and intervocalic intervals) (Grabe & Low, 2002). By now, many studies have applied the rhythm metrics to different languages and even non-native accents (Arvaniti, 2012; Dankovičová & Dellwo, 2007; Mok, 2009; Nolan & Asu, 2009; O'Rourke, 2008).

Criticisms of the rhythm metrics quickly followed, however, drawing evidence from problems in the computation, their instability due to speech rate, speaking style, within-speaker variation and measurement uncertainty, and their failure to clearly separate languages into the alleged rhythm classes (Arvaniti, 2012; Arvaniti & Rodriquez, 2013; Bertinetto & Bertini, 2008; Deterding, 2001; Gibbon, 2003; Knight, 2011; Nolan & Jeon, 2014, Dellwo, Leemann, & Kolly, 2015; White & Malisz, 2020). Arvaniti (2012), for instance, tested the performance of rhythm metrics ∆C, %V, PVIs, and Varcos in English, German, Greek, Italian, Korean and Spanish. The results suggest that the performance of the metrics depends largely on factors like inter-speaker variation, elicitation mentioned and the syllable composition of materials. Arvaniti and Rodriquez (2013, p.9) concluded that "values obtained using timing- based metrics cannot be seen as immutable properties of the languages involved but, rather, as points in wide distributions which overlap substantially across languages." This overlap is in terms of timing pattern and contradicts the view that languages belong to different rhythm classes with different timing features (cf. Loukina et al. 2011 for similar conclusions). Nolan and Jeon (2014) pointed out that rhythm metrics are not good at capturing rhythm classes and even if rhythm metrics do separate languages based on syllable-timing versus stress-timing to some degree, they do not provide conclusive evidence regarding whether the perceived differences are the result of either a deliberate rhythmic intention of the speaker or an inherent cyclicity pattern in speech production process.

It has also been suggested that rhythm is a perceptual phenomenon rather than a fact of speech production (Arvaniti, 2009; Kohler, 2009; Nakatani, O'Connor & Aston, 1981), and that "we hear speech as more regular than it physically is" (Eriksson, 1991, pp. 62). In support of this suggestion, Lehiste (1977) shows that durational differences smaller than 30 ms are never reliably identified and concludes that "sentences that are not produced with absolutely isochronous intervals between stresses may still be perceived as if the interstress intervals were identical." But as demonstrated by Nakatani et al. (1982), durational differences between inter-stress intervals can easily exceed 30 ms. What is critical for the rhythm class hypothesis is that it is not whether listeners hear something rhythmical in a language, but whether they can consistently determine whether a language is syllable-timed, stress-timed or mora-timed. This has been directly checked in Miller (1984) in which both trained phoneticians and naïve listeners are asked to classify languages as either stress-timed or syllable-timed. Not only is there no clear evidence that people have this ability, but also the classification by naïve listeners deviate from the rhythm class hypothesis more than trained phoneticians who are biased by the knowledge of the hypothesis. In other words, the perceptual impression of the phoneticians could be due to bias from their knowledge of the rhythm class hypothesis. Another test is done by Scott, Isard and de Boysson-Bardies (1985) who compared subjects' tapping to stimulus sentences in English and French. Their English subjects tapped regularly to units larger than syllables in French as much as they did English, while French subjects tapped regularly to French even more than the English ones. The finding has led to their conclusion that the perception-based tapping does not support the isochrony principle, or even the weak stress-timed/syllable-timed distinction. White, Mattys and Wiget (2012) investigated how language pairs were categorised by analysing utterances that only preserved durational features. They found that English listeners could distinguish between not only English and Spanish (from different rhythm classes), but also between different accents of British English. They believe that "considering the acoustic and perceptual evidence together, the categorical concept of rhythm class has little support" (White et al, 2012, p.677). Concerned that speaking rate might have been confounded with rhythm class in discrimination experiments, Arvaniti and Rodriquez (2013) modified speaking rate of their stimuli in their research and found that discrimination was possible both between and within rhythm classes when speaking rates differed between context and test. The participants in these discrimination experiments were adults. Additionally, there is evidence from infants. White, Delle Luche, and Floccia (2016) found that infants were able to distinguish French and Spanish (from same rhythm classes).

The perception findings demonstrate that languages may not be neatly classified in the way predicted by the rhythm class hypothesis and it is widely acknowledged among experts that there is no direct evidence of isochrony in any language. Consequently, the search for such evidence has been abandoned for a long time. However, the idea that languages are divided into rhythm classes based on timing is still widespread. This idea continues to drive rhythm-related research, and languages like English and Mandarin are still called stress-timed or syllable-timed.

1.4 Timing to mark major prosodic boundaries across languages

An important function of prosody is to provide cues for breaking up continuous speech into smaller chunks for ease of auditory comprehension. Of the variety of cues that have been reported, two are of particular importance, namely, pre-boundary lengthening and silent pause (Lehiste, 1972; Xu, 2009). Pre-boundary lengthening refers to the phenomenon that syllables and their component segments before a prosodic boundary are longer than they would be in other contexts (Klatt, 1975; Lehiste, 1972; Nakatani et al., 1981; Oller, 1973; Wightman et al., 1992; Xu & Wang, 2009). Also, the amount of pre-boundary lengthening is related to the strength of the boundary: the greater the strength, the longer the duration (Nakatani et al., 1981; Wightman et al., 1992; Xu & Wang, 2009). Silent pause, the second important boundary cue, is often associated with a strong boundary (Lea, 1980; O'Malley et al., 1973; Swerts, 1997). It is not yet clear, however, how exactly these two kinds of cues are distributed across boundaries of different strengths.

There has been some research on the relationship between boundary strength and pre-boundary lengthening. Wightman et al. (1992) demonstrated that pre-boundary lengthening varies significantly across all four levels of boundary strength: prosodic word, a group of words within a larger unit, intermediate phrase, and intonational phrase. Yang (1997) found that the duration of the syllable increases at first before word group and phrase boundaries, and then decreases before clause and sentence boundaries in Mandarin.

Some research has shown the relationship between silent pause and boundary strength. For Mandarin, it is found that normally there is no silent pause after prosodic word, but silence duration increases with the break level beyond prosodic word (Qian et al., 2001; Xiong, 2003; Yang & Wang, 2002). This seems to suggest a trade-off relation between pre-boundary lengthening and silent pause for larger boundaries in Mandarin. In English, 23% of the “intonation phrase” boundaries contained an unfilled pause, and 67% of the “groups of intonation phrases” contained an unfilled pause (Wightman et al., 1992). There have been suggestions that cues of lengthening and pausing may counterbalance each other (Lehiste, 1979; Scott, 1982).

Some research has proposed to combine pre-boundary duration and pause duration. This is based on the observation that both pre-boundary duration and pause duration affect the temporal distance between the onsets of the pre-boundary constituent and the post-boundary constituent (Xu & Wang, 2009). Xu and Wang (2009) then suggest using temporal distance to indicate relational distance between neighbouring constituents. This needs to be tested further.

1.5 Part of Speech effect

Part of Speech (POS hereafter) is taken as an essential input feature during training in both speech recognition and speech synthesis, but it is still not clear why it is important. POS may have an effect on duration. Assuming this effect does exist, it might be related to frequency of occurrence. As is known, verbs are more frequent than nouns and function words are more frequent than content words, so the Part of Speech effect and frequency effect are entangled.

Some researchers have performed experiments to investigate the POS effect on duration (Sorensen et al., 1973). They tested noun-verb homophones in sentences that were matched for phonetic environment and stress pattern and discovered that in typical sentences, nouns are longer than verbs. They did not, however, take grouping and frequency into account. In their sentences, the majority of nouns in their sentences are phrase-final, and verbs are phrase-initial. Furthermore, the frequencies of each noun and verb pair are different. As a result, the POS effect they found could be due to frequency and final lengthening. It is difficult to draw firm conclusions until all known or suspected factors have been clearly controlled.

Frequency of occurrence is already known to have an effect on duration. Reduced durations in high frequency morphologically complex Dutch words are observed by Pluymakers et al. (2005). In Mandarin, a newly mentioned word is slightly longer than a previously mentioned word

(Wang et al., 2017). Meunier and Espesser (2011) transformed lexical frequency to a 2-level factor and found that the dichotomized lexical frequency was significant on the vowel duration in CVC words, but not in CV words. Aylett and Turk (2004, 2006) measured language redundancy of a syllable by its predictability given its context and inherent frequency. They found a strong inverse relationship with syllable duration.

In this dissertation, I aim to provide a clearer link between the effects of POS and word frequency on duration than has previously been proposed. My hypothesis is that the POS effect on duration is actually derived from the effect of word frequency, and frequency has much greater predictive power than Part of Speech. This is performed by comparing phonetically identical Mandarin nouns and verbs.

1.6 Thesis goal and outline

The purpose of this thesis is to improve our understanding of timing by addressing the issues mentioned above. There are four chapters in the rest of the thesis. In Chapter 2, I revisit the rhythm class hypothesis by investigating compressibility of segments and syllables in English, an alleged stress-timed language, and Mandarin, an alleged syllable-timed language. Based on the observation of pre-boundary lengthening difference in Chapter 2, Chapter 3 compares temporal means to mark major prosodic boundaries in English and Mandarin. The results of Part of Speech effect and frequency effect are presented in Chapter 4. Finally, Chapter 5 concludes the main contributions of this thesis and suggests potential directions for future work.

2 Functional timing or rhythmical timing, or both? A corpus study of English and Mandarin duration

2.1 Background

Rhythm has received a lot of attention in the field of speech sciences. As mentioned in 1.3, previous research has classified languages of the world into three types: stress-timed, syllable-timed and mora-timed (Abercrombie, 1964a, 1964b, 1967; Pike, 1945; Ramus et al., 1999). According to this, in stress-timed languages, inter-stress intervals tend to be constant, and thus isochronous, whereas in a syllable-timed or mora-timed language, successive syllables or morae are equal in duration (Abercrombie, 1964, 1967; Pike, 1945). A great deal of experimental research has been carried out to search for evidence of such isochrony. However, no clear evidence has been found (Borzone de Manrique & Signorini, 1983; Lea, 1974; O'Connor, 1968; Pointon 1980; Roach,1982; Shen & Peterson, 1962; Uldall, 1971; Warner & Arai, 2001; Wenk & Wioland, 1981).

Some researchers try to reinterpret the rhythm class hypothesis with rhythm metrics (Dellwo, 2006; Dellwo & Wagner, 2003; Grabe & Low, 2002; Ramus et al., 1999) and the perceptual accounts of rhythm (Arvaniti, 2009; Kohler, 2009; Nakatani et al., 1981) by trying to bypass the core question about the hypothesis, namely, are there indeed any elements that tend to recur at a regular time interval as explicated by the rhythm class names? Ramus et al.’s (1999) metrics were proposed based on Dauer’s (1983) observation that stress-timed and syllable-timed languages have a number of different distinctive phonetic and phonological properties, of which the most important are syllable structure, vowel reduction and word stress. The proposed metrics therefore attempt to differentiate languages based on those phonological properties. This means that even if the parameters were able to separate languages into stress-timed and syllable timed languages as expected, they would have only validated the syllable structure and vowel reduction we already know about the languages examined but would have said nothing about whether there is isochrony at any level in those languages.

The rhythm class hypothesis has failed to find support in both production and perception studies as described in Section 1.3. Surface isochrony is not observed, but it does not rule out the possibility that there could be a top-down control of timing aiming at isochrony. According to van Santen & Shih (2000), if syllable duration does not increase by the same amount when there is some increment in the intrinsic duration of segments, there is compensatory effect. Such a compensatory effect could be interpreted as reflecting a tendency toward isochrony. It is important, however, to note that duration is also influenced by many other factors. Many such factors have indeed been recognized, some of which well before the proposal of the rhythm metrics (Ramus et al., 1999). These include, in particular, lexical stress (Fry, 1958; Klatt, 1976), focus (Bolinger, 1972;

Gussenhoven, 2008; Ladd, 1996; Turk & Shattuck-Hufnagel, 2000; van Heuven, 1994; Xu, 1999; Xu & Xu, 2005), boundary marking (Cutler et al., 1997; Lehiste, 1972), and intrinsic duration of segments (Klatt, 1976). The presence of these duration-affecting factors means that any tendency toward isochrony must be above and beyond their combined effects, and therefore the discovery of the isochrony tendency can only be made when these factors are systematically controlled.

Some research has already been done in this direction, although not always with the goal to search for evidence of isochrony. There are multiple findings that syllable duration in English increases quasi-linearly with syllable size, i.e., the number of constituent segments (Crystal & House, 1990; O'Connor, 1968; van Santen & Shih, 2000). Based on the database they examined, van Santen and Shih (2000) showed that syllable duration is highly predictable from segmental duration in English, i.e., with every increment in the intrinsic duration of segments, syllable duration increases by almost the same amount. One interpretation of this finding is that English syllable duration is not flexible enough to allow for any purely rhythm-driven timing control in the language. Interestingly, however, the authors found, in the same study, that syllable duration in Mandarin is not as highly correlated with vowel duration as in English. My interpretation, not contemplated in van Santen and Shih (2000), is that Mandarin syllable duration is more flexible than that of English, such that in Mandarin, syllable duration can indeed be described as showing a tendency toward rhythmic timing. I note, however, that the data is from one male speaker for English and one male speaker for Mandarin in their study, and thus, the generalizability of their findings is not yet clear. Along these lines, Nakatani, O'Connor and Aston (1981) find the duration of inter-stress intervals in English is at least linearly related to the number of constituent syllables, and that there is actually some accelerated increase of the interval duration with interval size. In that study, reiterant speech, whereby all syllables were replaced by [ma], was used to eliminate the segmental effects, which may have reduced the relevance of the findings to fully natural speech. But a similar linear relationship between the number of intervening unstressed syllables and the inter-stress interval for real words in sentence context was also found by Lea (1975) for English, although it was reported only in a conference abstract. These findings further suggest that English syllables are probably not compressed to maintain equal inter-stress intervals as the size of the inter-stress interval increases.

Nakatani et al. (1981) have also examined how syllable duration is affected by lexical stress and position in word and phrase in English. They find that both lexical stress and word/phrase position have clear effects on syllable duration, but the two kinds of effects work in parallel. For the positional effect, word-initial syllables are slightly longer than word-medial syllables, and interestingly, the duration of word-final syllables are roughly the same independent of whether the word is monosyllabic or polysyllabic. However, they did not provide statistical reports on these effects. Xu and Wang (2009), however, have found in Mandarin that phrase-medial syllables are shorter than phrase-initial syllables, and phrase-final syllables in multi-syllabic phrases are shorter than mono-syllabic words. Yuan and Liberman (2015) reported that word-medial plosives and affricates are more likely to be reduced than word-initial ones, which can be interpreted as a sign of shorter word-medial syllables than word-initial syllables. Compared to English, Mandarin therefore may have two additional means to shorten phrases as their sizes increase. One is to shorten phrase-medial syllables compared with phrase-initial syllables and the other is to shorten phrase final syllables from multisyllabic phrases compared with monosyllabic phrases. This makes it likely that Mandarin has a tendency toward isochrony of phrases, which, by the way, would run counter to the widely held belief that Mandarin is syllable-timed based on auditory impression, traditional analyses (Lin & Wang, 2007) and rhythm metrics (Grabe & Low, 2002; Lin & Wang, 2007; Mok & Dellwo, 2008).

With timing and isochrony back in focus, there therefore seems to be already evidence against the predictions of the rhythm class hypothesis in the case of English and Mandarin. English, an exemplary stress-timed language both by the original hypothesis and by the rhythm metrics, does not seem to have the flexibility of segmental and syllable duration needed to achieve stress timing. Mandarin, an alleged syllable-timed language by auditory impression and rhythm metrics, seems to exhibit signs of flexibility of segment and syllable durations that would permit tendencies toward both syllable timing and phrase timing. However, these findings have not led to a fundamental reconsideration of the rhythm class hypothesis, probably because the evidence is still rather scattered. There is a need to not only sort out the separate sources of evidence, as done in the above literature review, but also conduct a re-examination of the duration patterns of exemplary languages like English and Mandarin in close parallel, so as to test the generalizability of previous findings. For Mandarin, there is also a need to find out if the flexibility of segmental and syllable duration has indeed resulted in a tendency toward isochrony in units larger than the syllable.

This chapter presents a comparison of the timing patterns in two large non-experimental corpora, one in English and another in Mandarin, with the aim of both verifying previous findings from controlled experiments and answering further questions critical for the rhythm class hypothesis. The use of non-experimental corpora particularly allows the examination of units that are larger than those investigated previously. More specifically, the following questions are examined:

1) Is there a tendency toward isochrony in English?
    a. Are English segments adjustable toward equal syllable duration?
    b. Are English syllables adjustable toward equal inter-stress interval duration?
    c. Are English syllables adjustable toward equal phrase duration?
2) Is there a tendency toward isochrony in Mandarin?
    a. Are Mandarin segments adjustable toward equal syllable duration?
    b. Are Mandarin syllables adjustable toward equal phrase duration?

For 1a and 1b, I corroborate previous findings of linear relation between segment duration and syllable duration (Crystal & House, 1990; O'Connor, 1968; van Santen & Shih, 2000) in English, and between syllable duration and duration of interstress intervals in English (Bolinger, 1965; Lea, 1974; O'Connor, 1965; Shen & Peterson, 1962, Nakatani et al., 1981). For 1c, I examine whether there is a linear relation between syllable duration and phrase duration, where phrases may or may not coincide with interstress intervals. For 2a, I ascertain whether there is a linear relation between segment duration and syllable duration in Mandarin, just as in English, or whether segments are somewhat compressible to make syllables equally long. Previous findings on this, as mentioned above, have been equivocal (van Santen & Shih, 2000). Finally, for 2b, I will try to find out if in Mandarin, unlike in English, syllable duration is compressible to make it possible to approach equal duration of phrases. Previous findings by Xu and Wang (2009) have shown indications that this may be possible, as mentioned earlier.

2.2 Methods

2.2.1 English corpus

For English, the Boston University Radio News Corpus was used (Ostendorf, et al., 1995). It consists of news stories recorded by three female and four male FM radio news announcers during broadcast and the same four type-B news stories, which are normally pre-recorded and edited by announcers, recorded by six of the seven announcers in a laboratory condition. Number of sentences and average sentence length (number of words) for each speaker in the Boston University Radio News Corpus is shown in Table 1. The overall speech rate is 5.31 syllables per second. Professional radio announcers tend to be more fluent than non-professional speakers, producing fewer disfluencies and prosodic errors (Ostendorf et al., 1995). The paragraphs are annotated previously with orthographic transcriptions, phonetic alignments, part-of-speech tags and prosodic labels in the ToBI system (Ostendorf et al., 1995). The ToBI (tone and break indices) system marks prosodic phrasing, phrasal prominence and boundary tones. For lexical stress, only two levels are distinguished: stressed and unstressed. The phonetic alignments are generated automatically using constrained speech recognition (Kimball et al., 1992). Segmentation times and phone durations are provided in units of 10 milliseconds. Phonetic alignments for the news recorded in the laboratory were hand-corrected by the corpus developer, while those recorded during broadcast were not. In my analysis, data from one of the male speakers were excluded for not having prosodic information. All other announcers' data with complete segment, syllable, and prosodic information were used. The amount of data analysed is therefore greater than in other studies that also made used of this corpus (Choi et al., 2005; Sun, 2002).

Table 1. Number of sentences and average sentence length (number of words) for each speaker in the Boston University Radio News Corpus.

Speaker	Number of sentences	Average sentence length
f1a	159	25
f2b	535	23
f3a	119	24
m1b	190	29
m2b	176	24
m3b	164	26
m4b	549	25

One problem with the corpus was that words were divided into syllables based on a dictionary that combined MOBY and SRI dictionaries, which did not consider resyllabification (Gao & Xu, 2010; Kelso et al., 1986; de Jong, 2001). For example, the dictionary divided the word *decade* into “d eh+1 k” and “ey d”. In spoken English, speakers tend to say it as “d eh+1” and “k ey d”, so that “k” is an onset. Resyllabification was therefore performed based on the following rules: (1) within a word, if a coda is followed by a syllable beginning with a vowel, the coda is treated as the onset of the next syllable (Campbell & Isard, 1991); (2) between words, if a coda is followed by a syllable beginning with a vowel, and the break index (Beckman & Ayers, 1997) is 1 or 2 without silence, the coda is also treated as the onset of the next syllable.

2.2.2 Mandarin corpus

The Mandarin data were from Annotated Speech Corpus of Mandarin Discourse (ASCCD, Li et al., 2000), which was set up and recorded at Institute of Linguistics, Chinese Academy of Social Sciences. There are 18 discourses, each consisting of 300-500 syllables and several paragraphs. The corpus consists of read speech from five male and five female Beijing speakers who speak standard Mandarin (Li et al., 2000). There are 5 male and 5 female speakers. All of them read the same text. There are 242 sentences and the average sentence length is 27 words. Some of the speakers are teachers with a Phonetics background. The overall speech rate is 5.16 syllables per second. Four annotation tiers, including the syllable tier, initial and final (onset and rhyme) tier, break index tier and stress tier, were labelled (Li et al., 2000). In total, 41,673 CV syllables, 18,486 CVC syllables and 10,647 CGV syllables were analysed. The CGV structure is unique to Chinese, where G stands for the semivowel glide between onset and nucleus.

An advantage of both corpora is that they are already annotated with break index by the developers. This provides a level of objectivity in my data analysis, although the definitions of the break indices are not identical for the two languages, as will be explained.

2.2.3 Measurement

2.2.3.1 Syllables related to segments

To understand whether segments are compressed if their intrinsic duration is relatively long, I examined how closely syllable duration is correlated with the intrinsic duration of segments (estimated average duration), similar to what is investigated by van Santen and Shih (2000). To make my results comparable, I made my measurements in a similar way to theirs.

Suppose I analyse CV syllables that share the same context, with the same stress, the same structure in terms of number of segments and their order. Then the only difference between these syllables is their segmental makeup such as whether a syllable starts with a [t] or a [b]. Likewise, for CV syllables starting with the same consonant, the only difference would be whether the vowel is, e.g., [u] or [i]. van Santen and Shih (2000) have shown that, under these circumstances, syllable duration is highly predictable from segmental duration in English. Interestingly, however, the data in the same study showed that in Mandarin, syllable duration is not as highly correlated with vowel duration as in English. This language difference, however, is not elaborated in van Santen and Shih (2000). When extracting tokens for my analysis, for example, for the syllable “F-EY,” I extracted the following information: frequency, average duration, maximum and minimum duration of this syllable in my corpus, as well as the average duration, maximum duration and minimum duration of each segment in this syllable. All these information from all different syllables was listed in an Excel file and prepared for analysis using the method proposed by van Santen and Shih (2000).

Here is a summary of the method from van Santen and Shih (2000). DUR(*c*•) is the average duration of all CV syllables starting with consonant *c*; DUR(*c*|*c*•) is the duration of *c* averaged across all vowels; and D_{inherent}(*v*) denotes the inherent duration of a vowel. This method also applies to vowels. DUR(*v*•) is the average duration of all CV syllables ending with vowel v. DUR(*v*|*v*•) represents the average duration of *v* across all consonants. D_{inherent}(*c*) refers to the inherent duration of a consonant. van Santen and Shih (2000) assumed the vowels (v) ranged from1, …, V and consonants (c) from 1, …, C.

(1)

$$DUR(c\ \cdot) = DUR(c|c\ \cdot) - E_{compensatory}(c) + (\frac{1}{V})\sum_{v=1}^{v=V} D_{inherent}(v)$$

(2)

$$DUR(\cdot\ v) = DUR(v|\cdot v) - E_{compensatory}(v) + (\frac{1}{C})\sum_{c=1}^{c=C} D_{inherent}(c)$$

Equation (1) and (2) suggest that “if there is no compensatory timing [i.e., $E_{compensatory}(c)$ =0], then a graph (Figure 3) illustrating syllable duration [DUR(*c*•)] as a function of segmental duration [DUR(*c*|*c*•)] is a line with a slope of 1 and an intercept of

$$(\frac{1}{V})\sum_{v=1}^{v=V} D_{inherent}(v)$$

and likewise for vowel duration.” (van Santen and Shih, 2000, p. 1018)

van Santen and Shih (2000) also assumed that the degree of compensatory shortening caused by consonant c on a vowel is greater for consonants that are intrinsically longer. To illustrate, they assumed a linear relationship with slope (1-α) and intercept -β.

(3)

$$E_{compensatory}(c) = (1-\alpha)D_{inherent}(c) - \beta$$

Then,

(4)

$$DUR(c\bullet) = \alpha DUR(c \mid c\bullet) + (1/V)\sum_{v=1}^{v=V} D_{inherent}(v) + \beta$$

Equation (4) shows the compensation effect in a syllable, as it measures how much the duration of a consonant or vowel depends on the identities of the remaining segments in the syllable. The duration of the syllable as a function of segmental duration is illustrated in Figure 3, where α represents the slopes of the regression line. When α is 1, there is no compensation. When α is 0, there is complete compensation. Values of α between 0 and 1 indicate that there is partial compensation, hence, partial compression and/or elongation of segments in the direction of making syllables equally long.

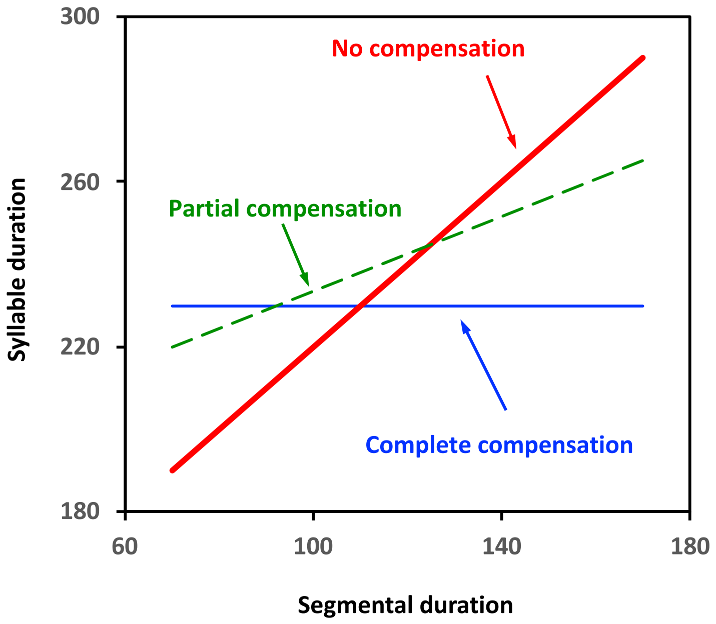


Figure 3. Schematic drawing of relation between syllable and segmental duration. It shows complete, partial, or no compensation, where “compensation” refers to how much the duration of a consonant or vowel depends on the identities of the remaining segments in the syllable (Adapted from van Santen & Shih, 2000).

Our investigation differs from van Santen and Shih (2000) in two ways, however. Firstly, they used an English database consisting of 2017 isolated sentences read by one American English male speaker and a subset of a database consisting of 424 Mandarin sentences recorded by one male Mandarin speaker. I used 369 paragraphs of news in English from three female speakers and three male speakers and a Mandarin corpus consisting of 18 discourses spoken by 10 speakers. Secondly, they reported only results of stressed word-initial CV syllables in phrase-medial words and stressed word-final CVC syllables in accented phrase-medial words in English, without considering consonant clusters. I treated consonant clusters as singletons and included CV and CVC syllables in all positions. More detailed differences are shown in Table 2.

Table 2. The differences between van Santern & Shih (2000) and current study.

		van Santern & Shih (2000)	Current study
Corpus	English	● 1 male speaker ● 2017 isolated sentences	● 3 female and 3 male speakers ● 369 paragraphs of news
	Mandarin	● 1 male speaker ● 1 subset of a database consists of 424 Mandarin sentences	● 5 female and 5 male speakers ● 18 discourses
Syllable structure	English	● CV and CVC ● Stressed syllables only ● No consonant clusters ● Certain positions	● CV and CVC ● All syllables ● Consonant clusters ● All positions

The underlying assumption is that the control of the effects of linguistic factors, i.e., lexical stress and boundary strength, is accomplished by using a very large number of syllables across all conditions to even out the influence, which may be compared to the use of long-term spectrum to examine speaker characteristics (Hollien and Majewski, 1977). An important reason for applying this control method is that intrinsic duration and syllable duration are both affected by stress and break index, and it would be difficult and unnecessary to separate their effects for the current purpose. Future research is needed to investigate this assumption more carefully.

2.2.3.2 Syllable duration in inter-stress intervals and phrases

We also examined whether and how closely syllable duration is related to linguistic factors of stress, and position in words/phrases; also, how inter-stress interval duration is related to number of syllables. Inter-stress interval is similar to the type of foot beginning with a stressed syllable followed by zero or more unstressed syllables, as described in Nakatani et al., 1981. Additionally, I investigated whether and how closely syllable duration is related to stress and positions in words and phrases. Here my method is similar to that of Nakatani et al. (1981), but with three major differences as shown in Table 3.

Table 3. Differences between Nakatani et al. (1981) and current study.

Nakatani et al. (1981)	**Current study**
Reiterant speech	Corpora of natural speech
Isolated sentences	Paragraphs and discourses
American English only	American English and Mandarin

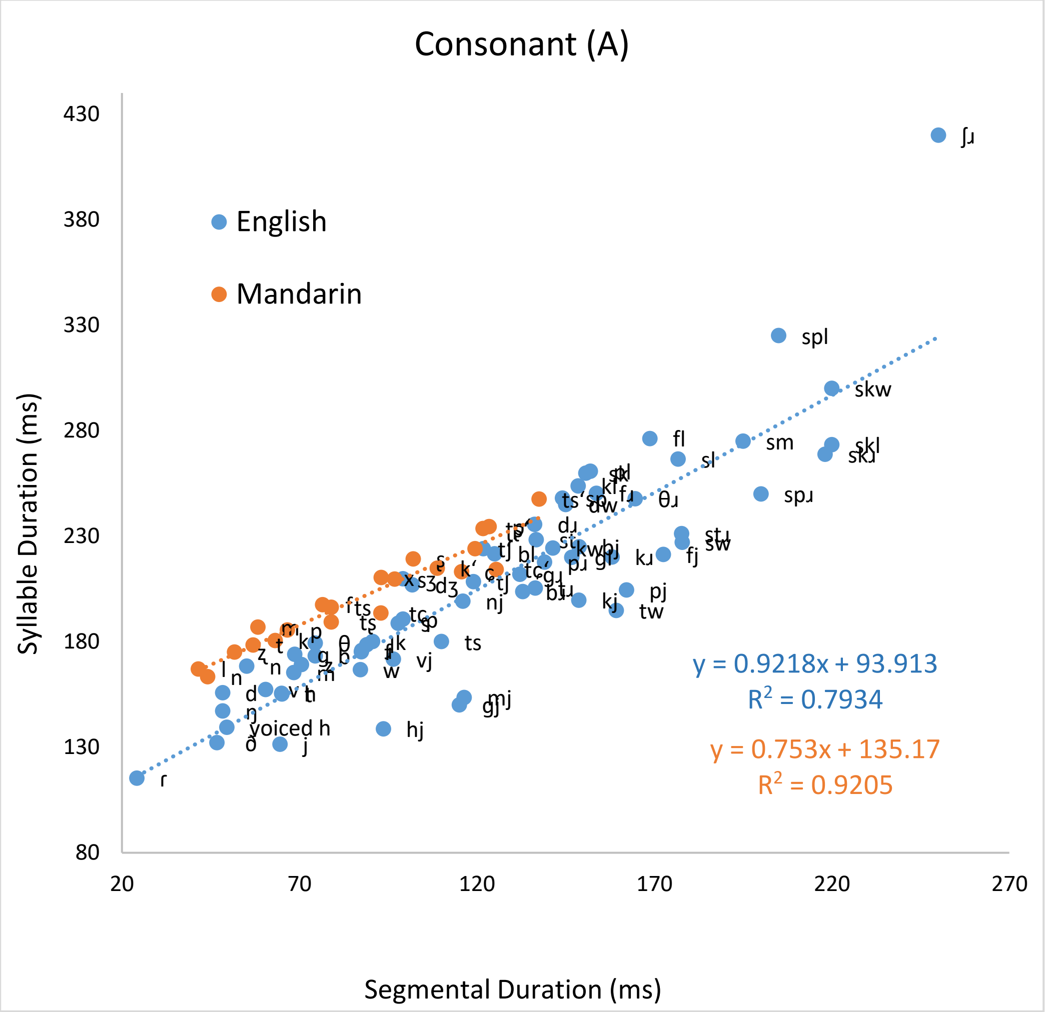
2.3 Results

2.3.1 Compressibility of segments

For syllables to show a tendency toward equal duration, their component segments must exhibit compressibility in one of two ways, or both. First, a segment would be compressed if its intrinsic duration is relatively long, so as to better match the intrinsically shorter ones. Second, all segments would be compressed as the number of segments increases in a syllable. In the following, I will examine both kinds of compressibility.

2.3.1.1 Relation of syllable duration to intrinsic segment duration

In this section, first I compare relation of CV syllable durations to intrinsic segment durations in American English and Mandarin. For illustration purpose, the phone labels in the English corpus, which are based on the TIMIT phonetic labelling system (Lyons, 1993), were replaced by IPA symbols in my analysis. The same procedure was done to phonetic labels in Mandarin which is based on Pinyin. Figure 4 shows plots of syllable duration as a function of intrinsic durations of onset and nucleus segments in CV syllables in American English (N = 18,941) and Mandarin (N = 41,673), and coefficients of Pearson correlation coefficients. For English, the coefficients are 0.891 (p < 0.001) and 0.936 (p < 0.001), and the slopes of regression lines are 0.9218 and 0.9736, respectively. For Mandarin, the Pearson correlation coefficients are 0.959 (p < 0.001) and 0.839 (p < 0.001), and the slopes of regression lines are 0.753 and 0.8131, respectively. In both languages, therefore, syllable duration is closely related to the intrinsic durations of the onset and the nucleus, but the slopes of regression lines are shallower in Mandarin than in English for both consonants and vowels.


Consonant (A)
English
Mandarin
Syllable Duration (ms)
Segmental Duration (ms)
430
380
330
280
230
180
130
80
20
70
120
170
220
270
ʃɹ
spl
skw
fl
sl
sm
skl
skɹ
spɹ
θɹ
stɹ
sw
fj
kɹ
pj
tw
kj
bj
kw
gl
pɹ
bɹ
gɹ
dɹ
st
sp
sk
sn
kl
fɹ
dw
ts
bl
tʃ
tɕ
nj
ts
mj
gj
hj
vj
w
dʒ
ʒ
x
ʃ
k
p
t
m
n
l
g
b
θ
z
v
h
d
ŋ
ð
j
voiced h
r
y = 0.9218x + 93.913
R^2 = 0.7934
y = 0.753x + 135.17
R^2 = 0.9205

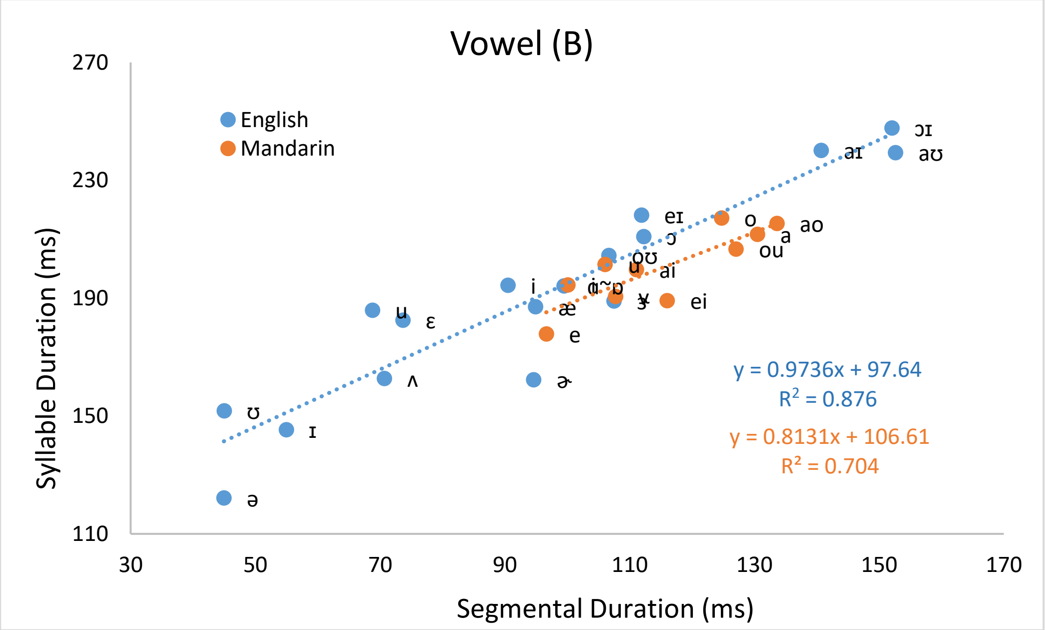
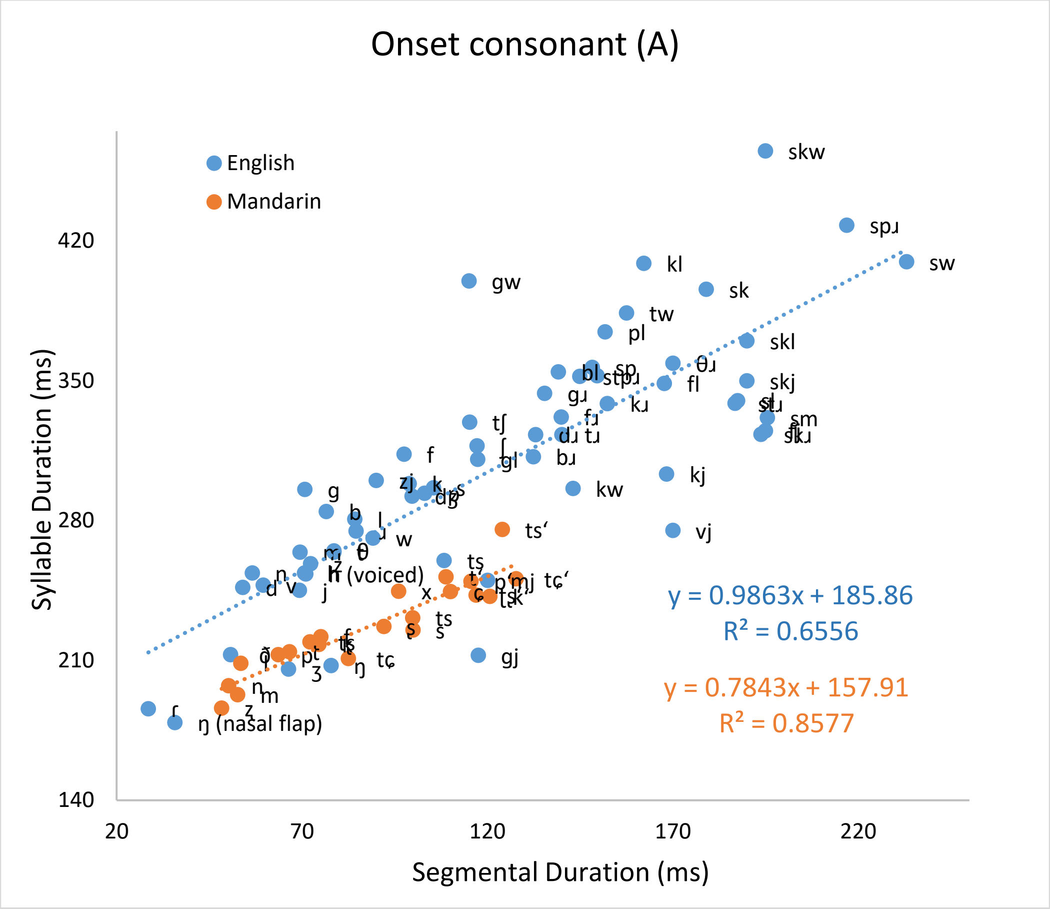
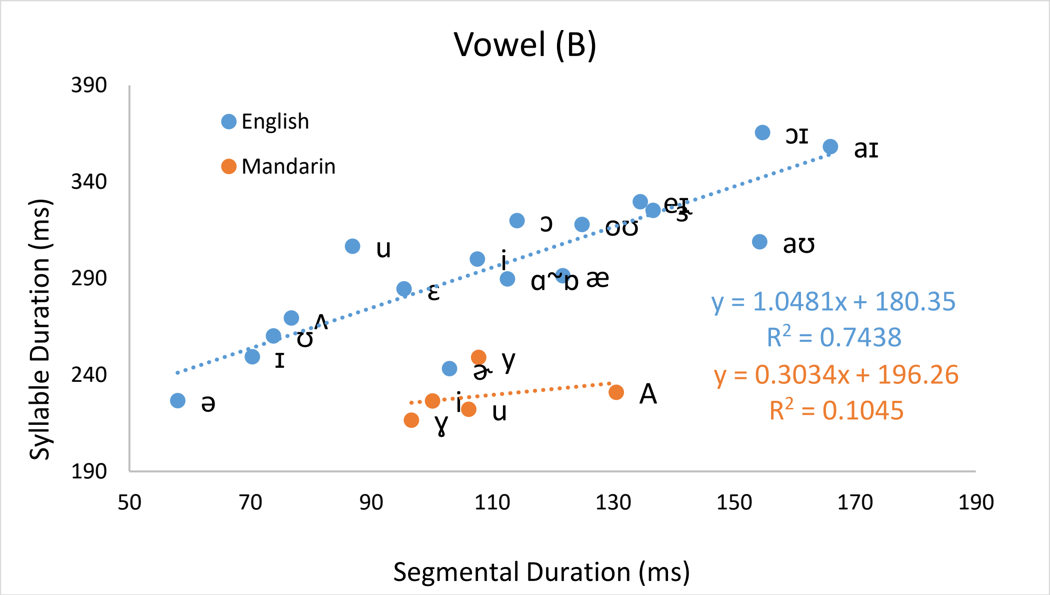

Vowel (B)
English
Mandarin
Syllable Duration (ms)
Segmental Duration (ms)
270
230
190
150
110
30
50
70
90
110
130
150
170
ɔɪ
aɪ
aʊ
eɪ
o
ao
a
ɔ
ou
oʊ
u
ai
i
ɑ
ɒ
ɤ
ɜ
ei
æ
e
u
ɛ
ʌ
ɚ
ʊ
ɪ
ə
y = 0.9736x + 97.64
R^2 = 0.876
y = 0.8131x + 106.61
R^2 = 0.704

Figure 4. Duration of CV syllable in American English and Mandarin as a function of intrinsic duration of consonants (A) and vowels (B), together with linear regression lines and Pearson correlation coefficients.

Next, I compare relation of CVC syllable durations to intrinsic segment durations in American English with that in Mandarin. Codas are not analysed in Mandarin, because there are only two codas, /n/ and /ŋ/, and they were not segmented in the corpus, so it was impossible to get their intrinsic durations. Figure 5 shows plots of syllable duration as a function of intrinsic durations of onset, nucleus and coda segments in CVC syllables in American English (N = 17,354) and Mandarin (N = 18,486), and Pearson correlation coefficients. For English, the correlations between syllable durations and segmental durations are 0.810 ($p < 0.001$) for the onset consonant, 0.862 ($p < 0.001$) for the vowel, and 0.815 ($p < 0.001$) for the coda consonant, and the slopes of the regression lines are 0.9863, 1.0481 and 1.0398, respectively. For Mandarin, the correlations between syllable duration and segmental durations are 0.926 ($p < 0.001$) for the onset, 0.323 for the vowel, and the slopes of regression lines are 0.7843 and 0.3034, respectively.


Onset consonant (A)
● English
● Mandarin
Syllable Duration (ms)
420
350
280
210
140
Segmental Duration (ms)
20
70
120
170
220
skw
spɹ
sw
kl
gw
sk
tw
pl
skl
bl
sp
stpɹ
θɹ
fl
gɹ
kɹ
skj
sl
stɹ
sm
fɹ
skɹ
fk
tʃ
dɹ
tɹ
ʃ
f
gl
bɹ
kj
zj
k
s
dʒ
kw
g
b
l
ɹ
w
vj
m
θ
z
h (voiced)
ts'
n
d
v
j
tʂ
t'
p'
mj
tɕ'
x
ɕ
tʂ
k'
ts
ʂ
s
f
tʂ
k
ð
p
t
l
ʒ
ŋ
tɕ
gj
n
m
ɾ
z
ŋ (nasal flap)
y = 0.9863x + 185.86
R² = 0.6556
y = 0.7843x + 157.91
R² = 0.8577


Vowel (B)
● English
● Mandarin
Syllable Duration (ms)
390
340
290
240
190
Segmental Duration (ms)
50
70
90
110
130
150
170
190
ɔɪ
aɪ
eɪ
ɜ
ɔ
oʊ
aʊ
u
i
ɑ
ɒ
æ
ɛ
ʌ
ʊ
ɪ
ɚ
y
ə
A
i
u
ɣ
y = 1.0481x + 180.35
R² = 0.7438
y = 0.3034x + 196.26
R² = 0.1045

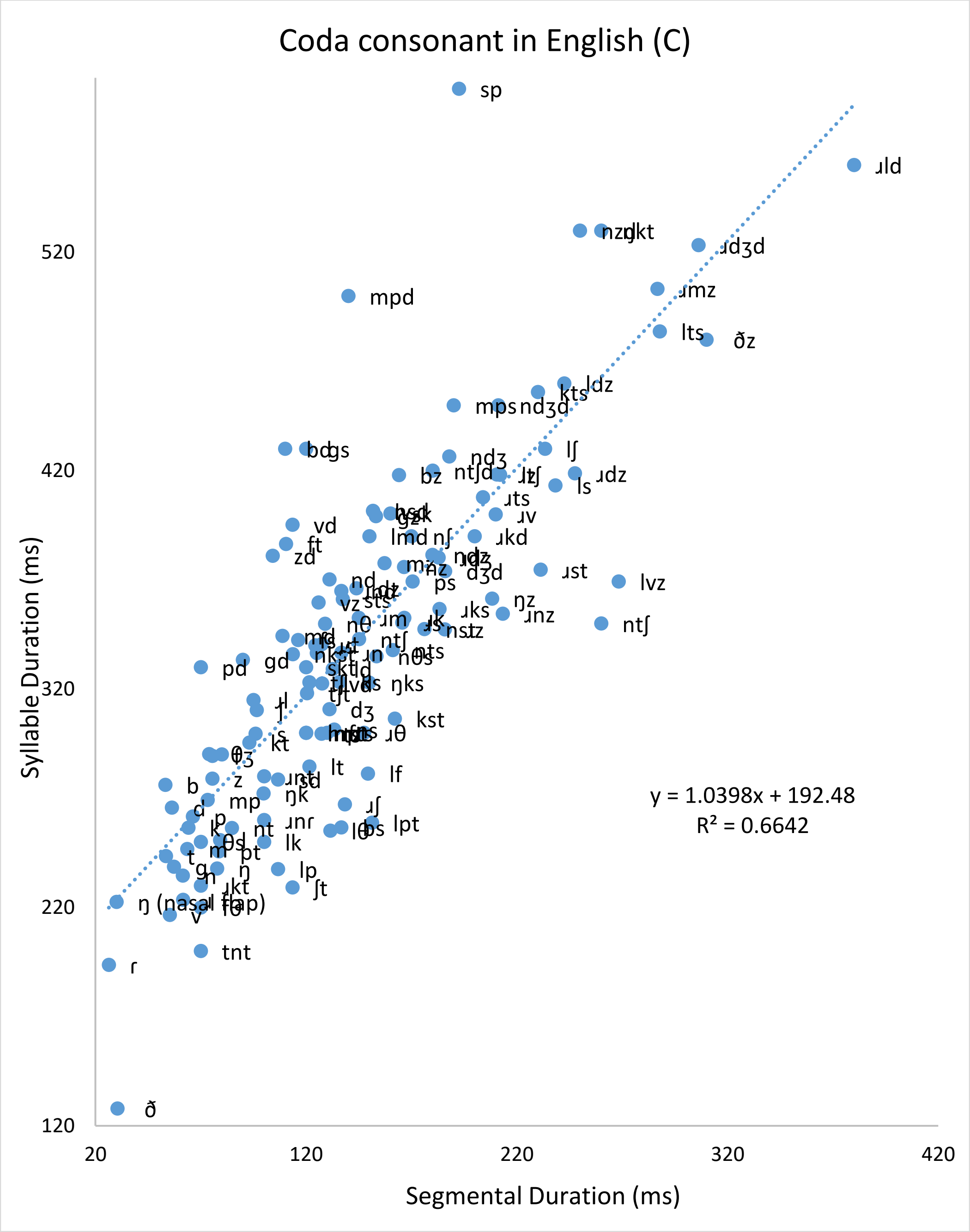
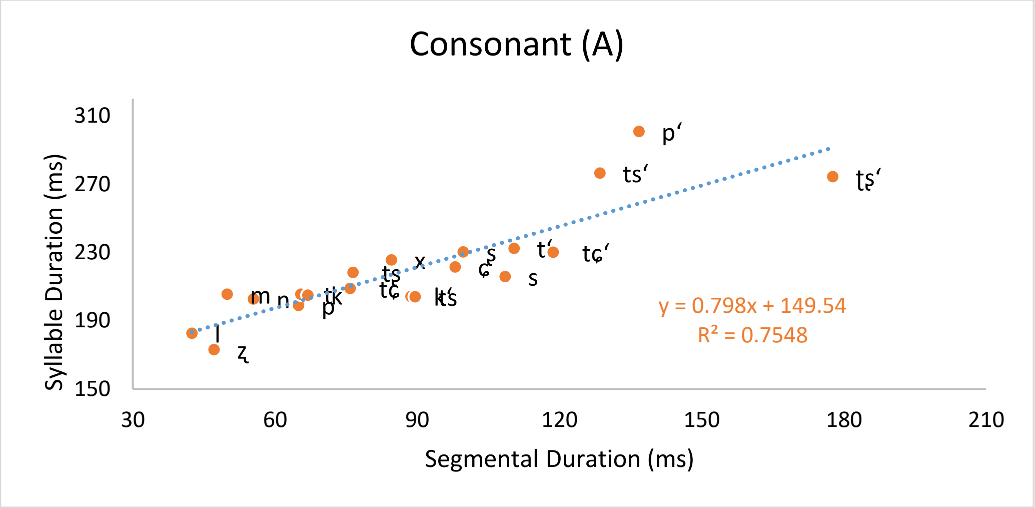
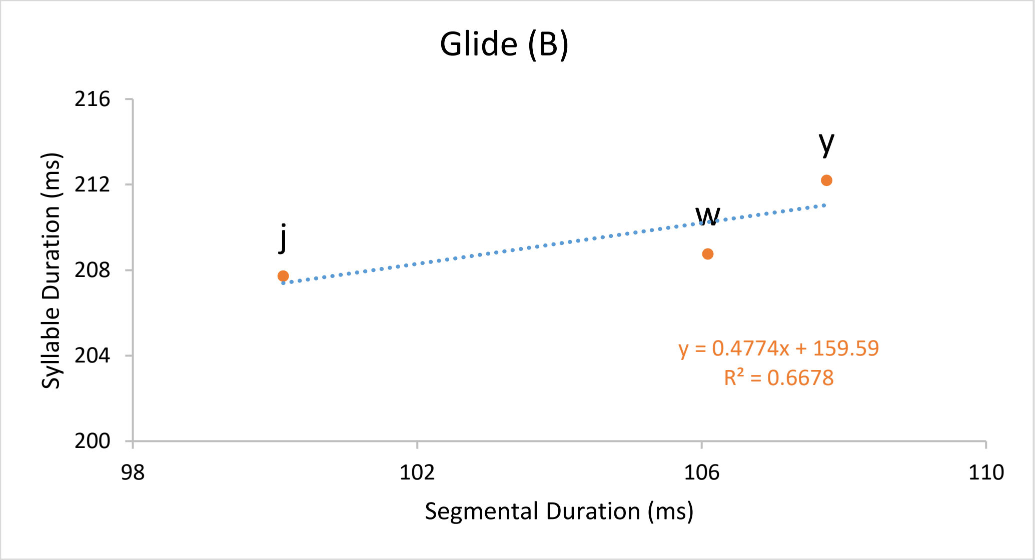


Figure 5. Effects of consonant (A and C) and vowel (B) identity on CVC syllable duration in American English and Mandarin, together with linear regression lines and Pearson correlation coefficients.

Next, I focus on CGV syllables in Mandarin, where G indicates a glide. Figure 6 shows plots of syllable duration as a function of intrinsic durations of consonant, glide and vowel in CGV syllables in Mandarin (N = 10,647), and Pearson correlation coefficients. The correlations between syllable duration and segmental duration are 0.869 (p < 0.001) for the onset consonants, 0.817 for the glides (p < 0.001) and 0.477 for the vowels (p = 0.279). The slopes of the regression lines are 0.798, 0.4774 and 0.4736, respectively.


Consonant (A)
Syllable Duration (ms)
310
270
230
190
150
30
60
90
120
150
180
210
Segmental Duration (ms)
pʻ
tsʻ
tʂʻ
tʻ
tɕʻ
x
ʂ
ɕ
s
tʂ
tɕ
kʻ
ts
m
n
t
k
p
l
ʐ
y = 0.798x + 149.54
R² = 0.7548


Glide (B)
Syllable Duration (ms)
216
212
208
204
200
98
102
106
110
Segmental Duration (ms)
j
w
y
y = 0.4774x + 159.59
R² = 0.6678

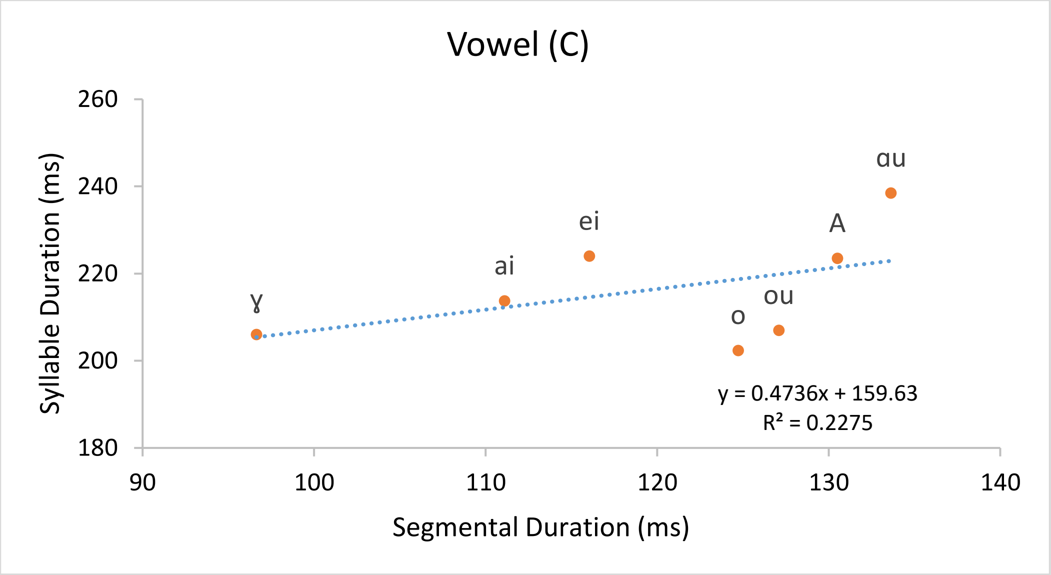
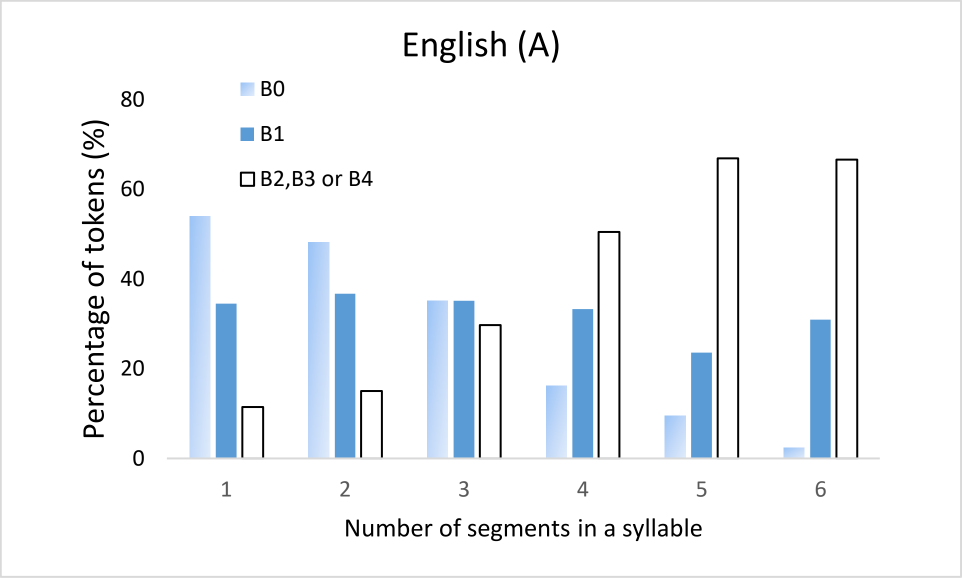

Vowel (C)
Syllable Duration (ms)
260
240
220
200
180
90
100
110
120
130
140
Segmental Duration (ms)
ɤ
ai
ei
A
ɑu
o
ou
y = 0.4736x + 159.63
R² = 0.2275

Figure 6. Effects of onset consonant (A), glide (B) and vowel (C) identity on CGV syllable duration for Mandarin, together with linear regression lines and Pearson correlation coefficients. The most important result shown so far is that the slopes of the regression lines for syllable duration as a function of intrinsic duration of segments are close to 1 for English in both CV and CVC syllables. These results can be interpreted as showing that compared to Mandarin, English segments maintain their intrinsic durations; the segments are neither compressed nor stretched to make syllables equal in duration. In contrast, for Mandarin, the slopes of the regression lines are well below 1.0 in both CV and CVC syllables, especially in the latter. The slopes are especially shallow for vowels and glides. This suggests that Mandarin has a tendency to adjust segment duration in order to maintain a constant syllable duration.

2.3.1.2 Relation of syllable duration to syllable size

In this section, I examine only the syllables that occur before a B1 boundary. A B1 boundary refers to most phrase-medial word boundaries in English (Beckman and Ayers, 1997) and prosodic word boundary (Li, 2002); Hence, these syllables never occur in final phrase/final utterance position.For the relation of syllable duration and syllable size (number of component segments), a potential confounding factor is that, in English, there is an uneven distribution of syllables of different sizes across boundaries of various strengths. Figure 7 shows histograms of syllables of various sizes at different boundary indices. As can be seen, 53.98% of the one-segment syllables occur before B0, while 66.67 % of the six-segment syllables occur before a phrase boundary (B2, B3 and B4). In contrast, syllables of different sizes are much more evenly distributed before B1. Although the same trend is not seen in Mandarin, to avoid the potential bias it may bring, in the following analysis, I thus include only syllables before B1 in both English and Mandarin. Also excluded from the analysis are syllables with the neutral tone in Mandarin.



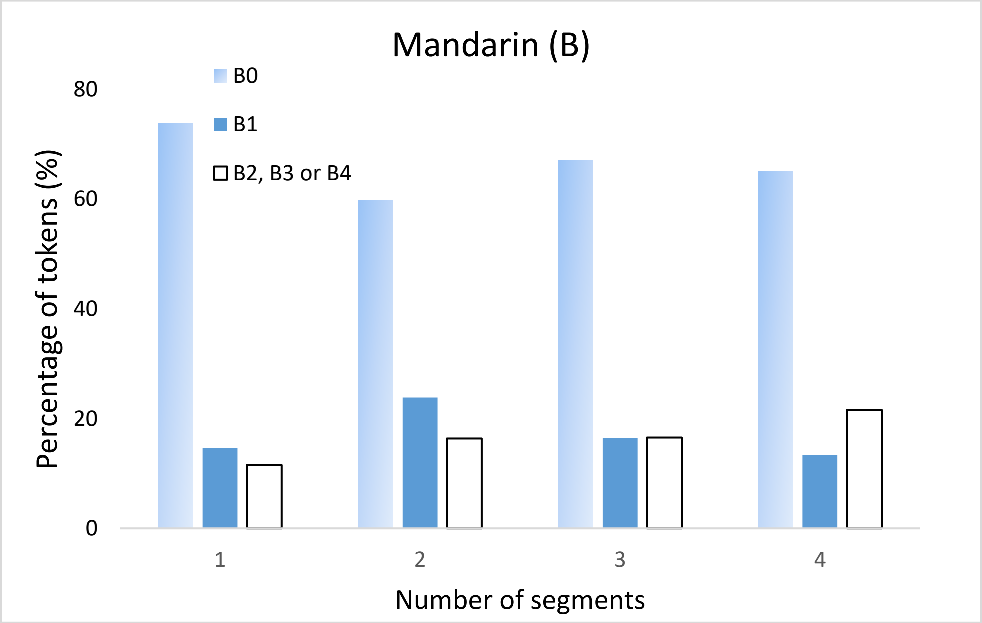
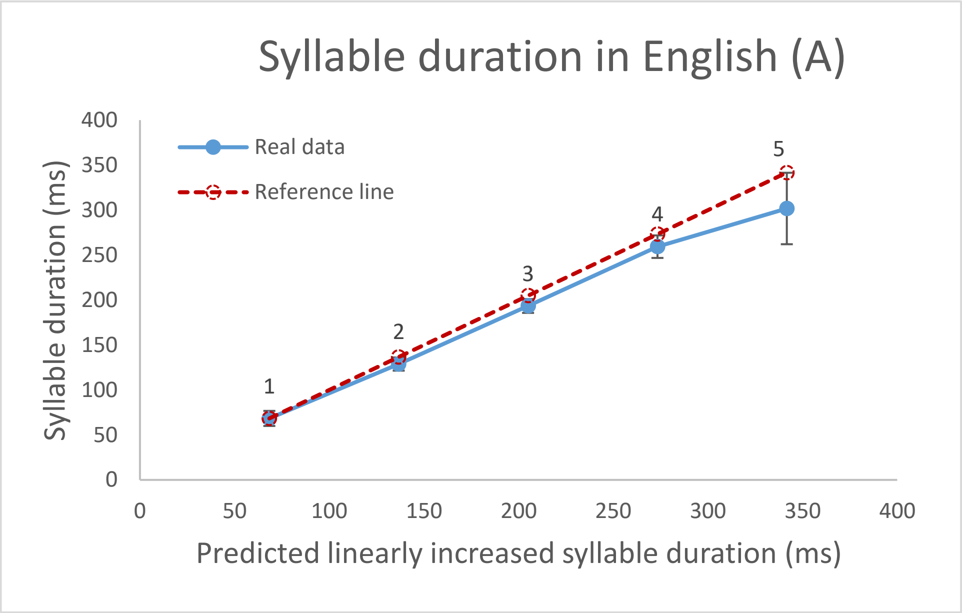
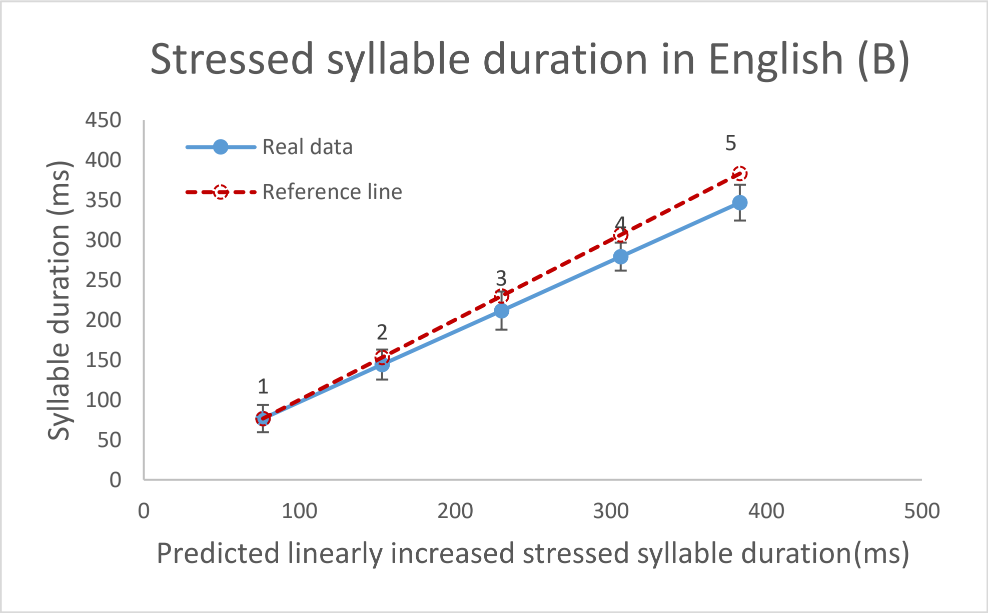


Figure 7. Histograms of distribution (% of tokens) of syllables of different sizes before different levels of boundaries (x axis) in English and Mandarin.

Figure 8 shows syllable duration in English and Mandarin as compared to the linear reference (dashed) lines. Because stress plays a role in the relation between syllable size and syllable duration, especially in English, the results from stressed and unstressed syllables are presented separately. The solid lines were drawn from syllables consisting of 1-5 segments in English and 1-4 segments in Mandarin. The dashed lines (reference lines) refer to predicted syllable duration by treating syllables with only one segment (leftmost point) as the reference. In current study, the average duration of syllables with only one segment is the sum of duration of all syllables with only one segment divided by the number these syllables in each corpus (English and Mandarin). Based on these calculations, the average duration of syllables with just one segment is 68 ms in English, 77 ms in stressed syllable duration and 65 ms in unstressed syllable duration in English and 146 ms in Mandarin if there is no tendency of equal duration. Treating syllables with only one segment as the reference means that each additional segment in the syllable should increase syllable duration by 68 ms in English syllable duration, 77 ms in stressed syllable duration and 65 ms in unstressed syllable duration in English and 146 ms in Mandarin, assuming there is no tendency of equal duration. As can be seen, as the number of segments increases, syllable duration increases almost linearly in English, although the rate of increase is reduced slightly in the most complex syllables (those consisting of five segments). In Mandarin, in contrast, the rate is substantially reduced starting from 2-segment syllables.


Syllable duration in English (A)
Real data
Reference line
Syllable duration (ms)
400
350
300
250
200
150
100
50
0
1
2
3
4
5
0
50
100
150
200
250
300
350
400
Predicted linearly increased syllable duration (ms)


Stressed syllable duration in English (B)
Real data
Reference line
Syllable duration (ms)
450
400
350
300
250
200
150
100
50
0
1
2
3
4
5
0
100
200
300
400
500
Predicted linearly increased stressed syllable duration(ms)

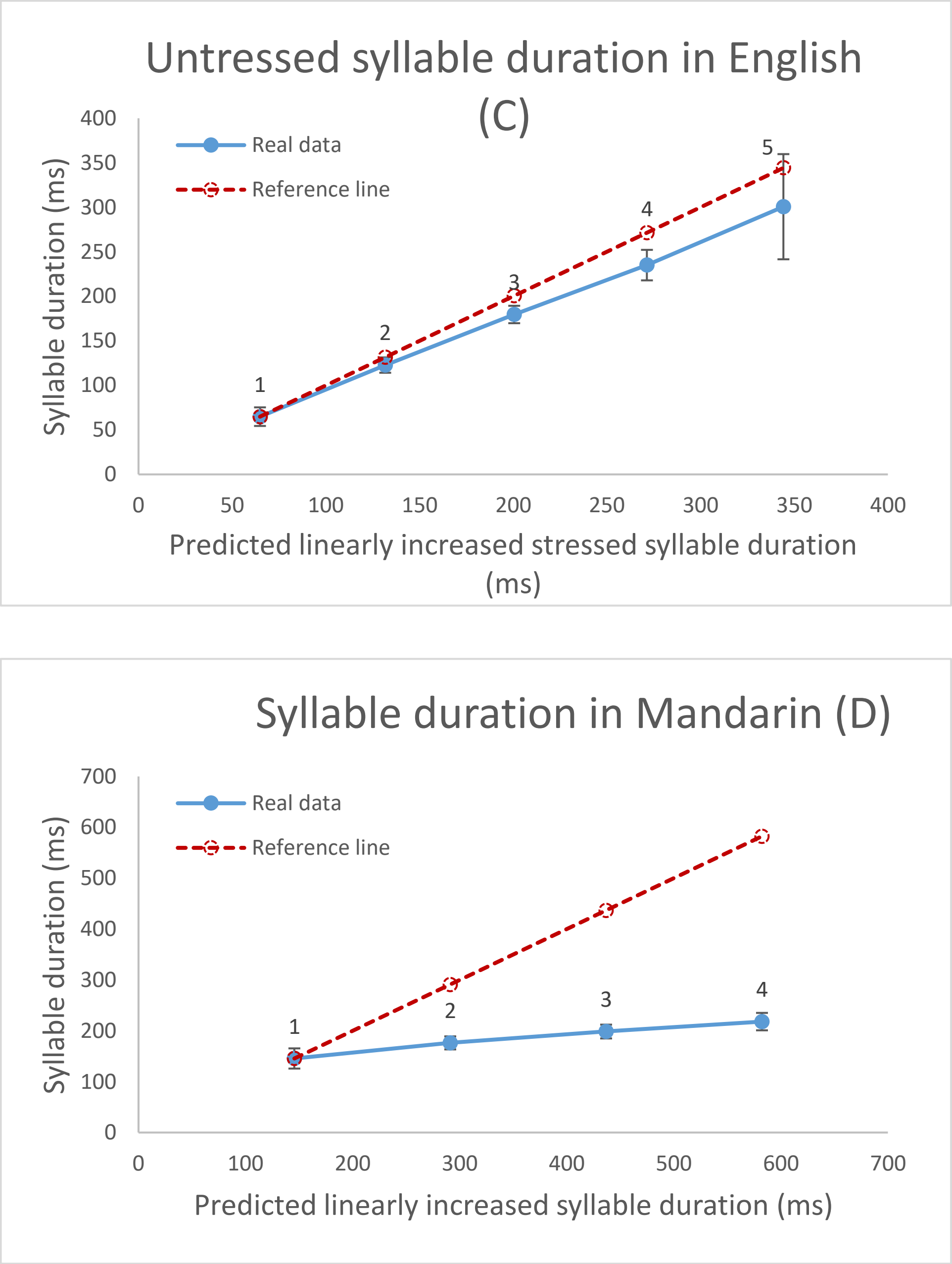


Figure 8. Mean syllable duration in English and Mandarin as a function of syllable size (number of component segments) for all English syllables (A), English stressed syllables (B) English unstressed syllables (C) and Mandarin syllables (D). The number of syllables is indicated by a number above each point in the graph.

The reduction of rate of increase in syllable duration as a function of syllable size in English (Figures 8a-c) occurs mainly in syllables consisting of three, four and five segments. The source of this reduction is likely consonant clusters, as shown in Figures 9-10. 1028 onset consonant clusters and 2151 coda consonant clusters pooled from all speakers were included in the following analysis. Figure 9 and 10 display durations of each consonant in different locations in a cluster as compared to its intrinsic duration from CV syllables or CVC syllables.

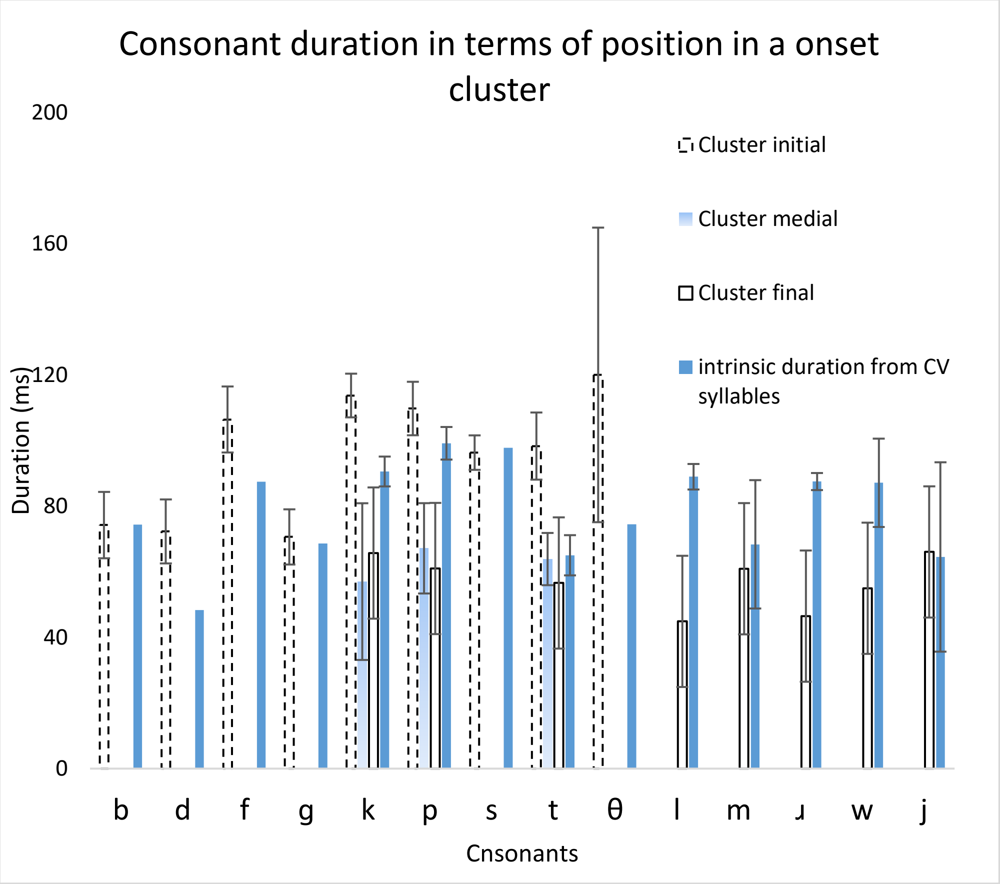


Figure 9. Consonant duration in initial, medial, final position in onset consonant clusters or as singletons, compared to the intrinsic durations of the same consonants calculated from CV syllables. CCCVC, CCCVCC, CCVC, CCVCC, CCVCCC syllables are pooled together.

Consonant duration at different within-cluster locations were compared with their intrinsic duration by Paired samples T-tests. Initial consonants (M = 9.57, SD = 1.89) are significantly longer than their intrinsic durations (M = 7.84, SD = 1.68); *t* (8) = -3.227, *p* = .012, n = 9, while final consonants (M = 5.71, SD = 0.80) are significantly shorter than their intrinsic durations (M = 8.14, SD = 1.34); *t* (7) = 3.954, *p* = .006, n = 8.Although there is a trend that medial consonants (M = 6.27, SD = 0.52) are shorter than their intrinsic durations (M = 8.49, SD = 1.78), there is no statistical significance; *t* (2) = 2.108, *p* = .170, n = 3.

Figure 10 shows intrinsic consonant duration and their duration in different locations within a coda consonant cluster. Paired samples T-tests show that here is no significant difference between initial consonants (M = 7.04, SD = 2.28) and their intrinsic durations (M = 7.30, SD = 2.32); *t* (10) = .958, *p* = .361, n = 11. Medial consonants (M = 4.15, SD = 1.48) are significantly shorter than their intrinsic durations (M = 7.78, SD = 3.01); *t* (5) = 5.171, *p* = .004, n = 6. Final consonant (M = 6.60, SD = 3.14) are also significantly shorter than their intrinsic durations (M = 7.78, SD = 3.01); *t* (5) = 6.227, *p* = .002, n = 6.

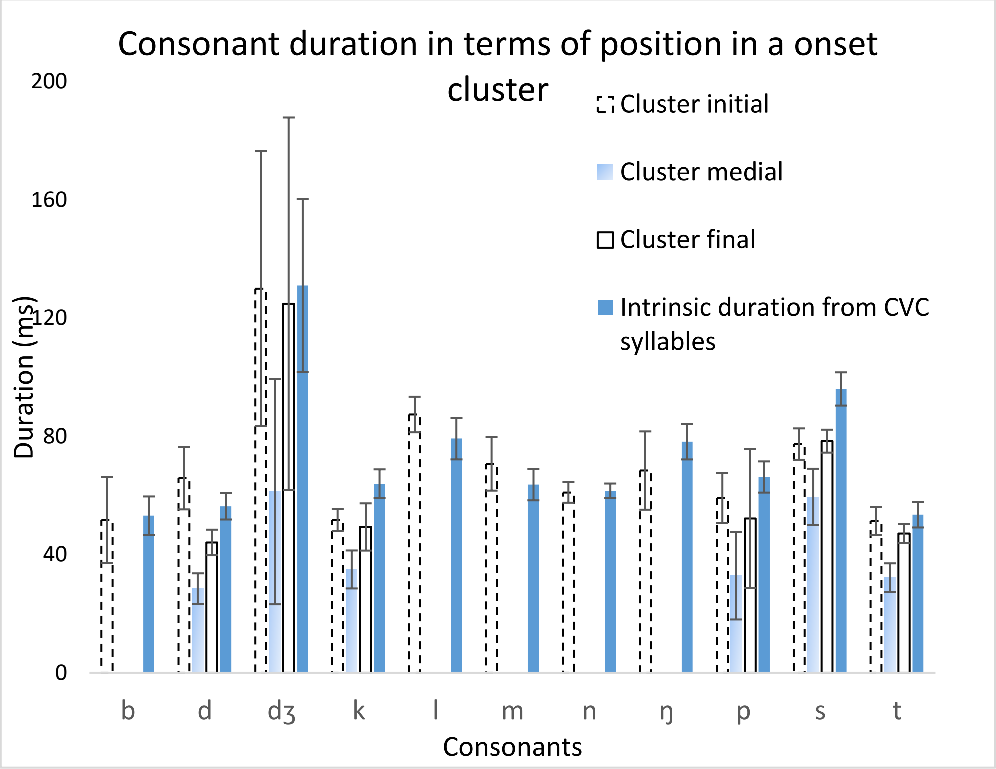


Figure 10. Consonant duration in initial, medial, final position in coda consonant clusters and intrinsic durations of the same consonants in coda, calculated from CVC syllables. CCCVCC, CCVCC, CCVCCC, CVCC, CVCCC syllables are pooled together here.

To summarize, despite a lengthening effect on initial consonants in onset clusters, there are significant shortening effects on final consonants in onsets, and on medial and final consonants in codas. Compression of consonant clusters may therefore be a main source of shortening in syllables consisting of five or more segments in English.

2.3.2 Compressibility of syllables

The results reported in section 2.3.1 show that English segments are not compressible for the sake of equal syllable duration, which contrasts with Mandarin where segments are clearly compressible in the direction of making syllables of different sizes equally long. This seems to be consistent with the classification of Mandarin as a syllable-timed language and English as a non-syllable-timed language. But it also leaves open whether either of the two languages shows a tendency toward equal timing at a level above the syllable. In the following analyses, I will examine for English if there is any tendency toward equal inter-stress intervals, and if there is any tendency toward equal phrase duration for both English and Mandarin.

2.3.2.1 Inter-stress intervals in English

According to the rhythm class hypothesis, inter-stress intervals are constant in a stress-timed language. If this is true, inter-stress intervals should maintain a constant duration regardless of the number of syllables in an interval, or at least show a tendency in that direction. Inter-stress intervals that were not phrase-final and immediately followed by a pause were included in the analysis (Nakatani et al., 1981). As shown in Figure 11, the average duration of unstressed syllable from "su" inter-stress intervals is treated as a reference for unstressed syllables. If the "su" inter-stressed intervals are phrase-final, the unstressed syllables are subject to phrase-final lengthening. Using their mean duration as the reference therefore is problematic, as phrase final lengthening does not apply to every syllable. To assess the relationship between the number of syllables and inter-stress interval duration, I measured from the onset of a stressed syllable to the onset of the next stressed syllable (Roach, 1982). I only considered intervals with one to four syllables, because each have more than 30 tokens from each speaker in my data. Longer intervals have too few tokens to guarantee reliability. In total, 7184 inter-stress intervals were analysed.

Figure 11 shows the average durations of inter-stress intervals as a function of size in terms of number of constituent syllables. Inter-stress interval duration is highly related to interval size. The correlation between inter-stress interval duration and interval size is 0.981 (p < 0.001). It is important to note that this result is not new. It is a confirmation of previous findings (Classe, 1939; Bolinger, 1965; Lea, 1974; O'Connor, 1965; Shen & Peterson, 1962). Contrary to the prediction of English being a stress-timed language and that the inter-stress intervals are constant regardless of number of syllables, the inter-stress intervals are linearly related to the number of syllables. As is shown in Figure 11, according to the reference line, each additional unstressed syllable should increase the inter-stress interval duration by 155 ms if there is no tendency towards equal inter-stress interval duration. However, the duration difference between the “s” and “suuu” inter-stress intervals is more than three times of the average duration of unstressed syllables (155 ms). This is similar to the acceleration found by Nakatani et al. (1981), although they excluded intervals with inter-stress function words. Speculatively, as the size increases, a stress group is more and more likely to break up into sub-intervals, and the boundary of the sub-intervals are marked by final lengthening, which would in turn increase the duration of the inter-stress interval as a whole. This possibility has also been raised by Uldall (1971, 1978). But systematic studies are needed in the future to examine it in depth.

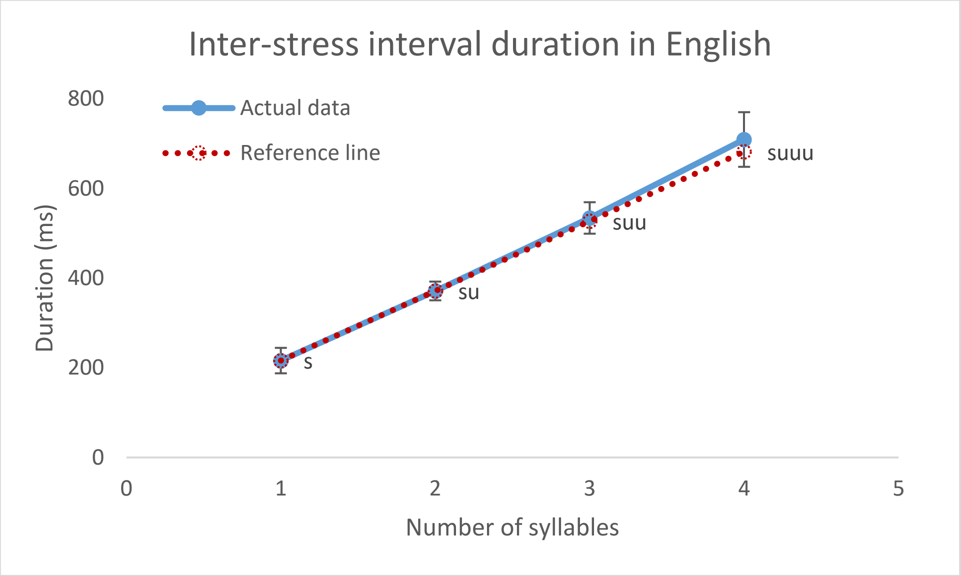


Figure 11. Average inter-stress interval durations in ms, the letter “s” indicates a stressed syllable and “u” indicates an unstressed syllable. On the reference line, the leftmost point shows the average duration of monosyllabic stressed syllables. The second left point on this line shows the average duration of "su" inters-stress intervals. The duration difference between "s" and "su" is the average duration of the unstressed syllables in "su" and it is 155 ms calculated from the English corpus. Then this duration difference is treated as a reference for unstressed syllables. This means that each additional unstressed syllable should increase inter-stress interval duration by 155 ms if there is no tendency towards equal inter-stress interval duration.

2.3.2.2 Compressibility of syllables in prosodic phrases in Mandarin and English

2.3.2.2.1 Phrase duration and phrase size

Mandarin does not have lexical stress that is equivalent to word stress in English. Even though the neutral tone shows phonetic properties similar to those of English unstressed syllables (Xu and Xu, 2005; Chen and Xu, 2006), its occurrence is infrequent. This makes it impossible to compare inter-stress intervals between the two languages. The two languages can be compared, however, in terms of phrase duration. This section examines for English and Mandarin whether syllables are compressed as the number of syllables in a phrase increases. Here I extracted prosodic phrases based on the break indices at or above B2 in the two corpora, i.e., B2, B3, and B4. In the English corpus, B2 refers to a lower-level perceived grouping of words without an intermediate B3 or a B4 intonational boundary marker or the break with next word is weaker than expected although the pitch pattern clearly suggests an intermediate B3 or a B4 intonation phrase boundary (see Beckman and Ayers, 1997). In Mandarin, B2 refers to minor prosodic phrase boundary; B3 refers to major prosodic phrase boundary; and B4 refers to prosodic group boundary (Li, 2002). Due to the difference in definitions, therefore, the durational relation of syllable to phrase can be compared only in terms of trends.

For English, lexical stress is a confounding factor, because each word can have only one primary stress and stressed syllables are much longer than unstressed syllables (Crystal & House, 1988; Nakatani et al., 1981). Thus, an increase in word length is necessarily achieved by adding more unstressed syllables, which may generate the appearance of a tendency toward isochrony. To control for stress, the duration difference of each segment or consonant cluster between stressed and unstressed syllables was calculated and added to every segment and consonant cluster in an unstressed syllable when computing phrase duration.

The first step was to calculate the differences between stressed segments and unstressed segments. Take [ei] as an example, stressed [ei] duration was calculated as an average duration of [ei] from all CV stressed syllables ending with [ei]. Unstressed [ei] duration was calculated as an average duration of [ei] from all CV unstressed syllables ending with [ei]. The difference between averaged stressed [ei] duration and averaged unstressed [ei] duration is the vowel duration difference for [ei]. This method was applied to each vowel. Those vowels that do not have both stressed and unstressed data were given a default duration difference which is the average duration difference of all other vowels that have data in both conditions.

Consonants were divided into onset and coda, and consonant clusters were treated as singletons. Take [p] for example. onset [p] duration in stressed condition was calculated as an average duration of [p] from all CV stressed syllables starting with [p]. Onset [p] duration in unstressed condition was calculated as an average duration of [p] from all CV unstressed syllables starting with [p]. The duration difference between averaged [p] in stressed condition and in unstressed condition is the duration difference for [p]. This method was applied to each onset consonant. Those onset consonants that do not have data in both stressed and unstressed conditions were given a default duration difference which is the average duration difference of all other onset consonants (excluding consonant clusters). Those consonant clusters that do not have data in both stressed and unstressed conditions were given a default duration difference which is the average duration difference of all other onset consonant clusters that have the same number of segments. A similar method was applied to coda consonant, in which the only difference is that the calculation was done on CVC syllables.

With this method, the shorter duration of unstressed syllables was not attributed to the reduction of phrase duration. As mentioned earlier, there is no lexical stress in Mandarin that is similar to word stress in English. Even if the neutral tone shows phonetic properties similar to those of English unstressed syllables (Xu and Xu, 2005; Chen and Xu, 2006), its occurrence is infrequent. For this reason, Mandarin phrases with one or more neutral tone syllables were excluded from the analysis. Figure 12 shows phrase duration in English (N = 6,523) and Mandarin (N = 7,406) as a function of phrase size in comparison with predicted linearly increased phrase duration. In both languages, phrase duration is strongly related to phrase size: Pearson correlation coefficients are 0.984 ($p < 0.001$) in Mandarin and 0.987 ($p < 0.001$) in English.

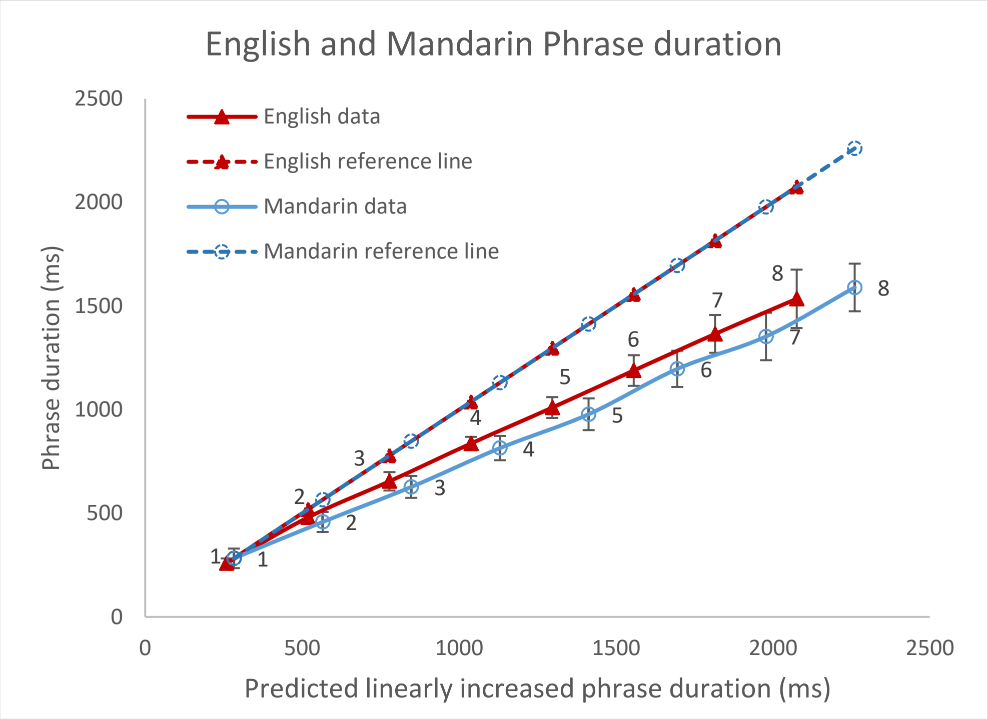


Figure 12. Average duration of phrases in English and Mandarin as a function of number of syllables in each phrase, in comparison with linearly predicted phrase duration. The solid lines were drawn from phrases consisting of 1-8 syllables. The dashed lines (reference lines) refer to predicted phrase duration by treating monosyllabic phrase as the reference. As a result of these calculations, each additional syllable in the phrase is supposed to increase phrase duration by 259 ms in English and 283 ms in Mandarin if there is no tendency of equal duration.

Compression can be seen in both languages from the graph. But this is likely because syllables in monosyllabic phrases, by definition, are phrase final, which is subject to phrase-final lengthening. Using their mean duration as the baseline therefore provides an inflated reference slope, as phrase final lengthening does not apply to every syllable. To circumvent this problem, I then examined the compressibility of words as the number of syllables in a word increases. A word is defined as one that is marked by break index 1 in both the English and Mandarin corpora.

2.3.2.2.2 Word duration and word size

Figure 13 shows word duration in English (N = 12,931) and Mandarin (N = 9,955) as a function of word size as compared with predicted linearly increased word duration. As can be seen, word duration is strongly related to word size in both language: Pearson correlation coefficients are 0.98 (p < 0.001) in Mandarin and 0.989 (p < 0.001) in English. But it can be also seen that syllables are compressed more in Mandarin than in English.

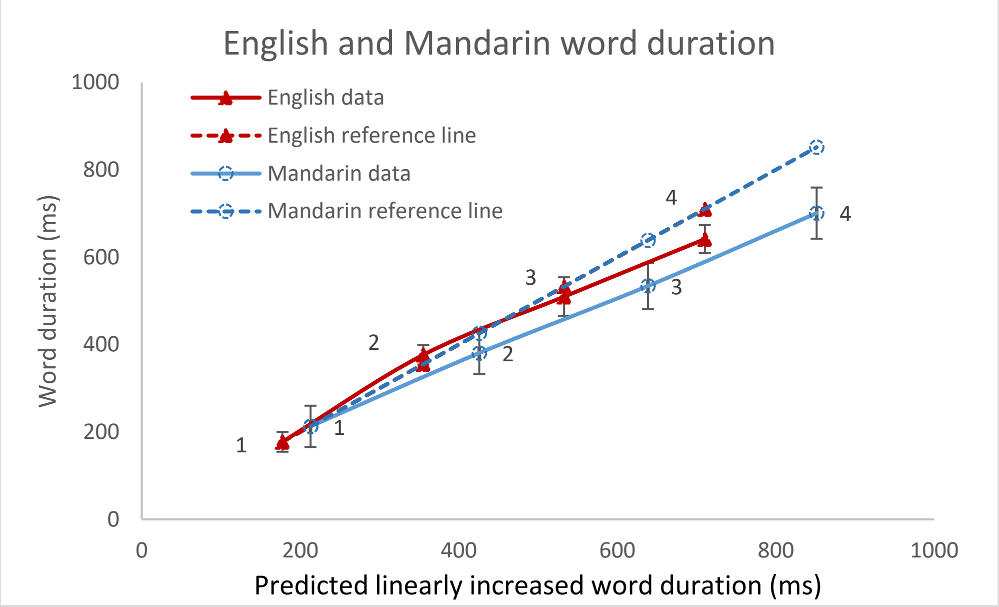


Figure 13. Word duration in English and Mandarin as a function of predicted linearly increased word duration. The solid lines were drawn from words consisting of 1-4 syllables. The dashed lines (reference lines) refer to predicted word duration by treating monosyllabic word duration as the reference. As a result, each additional syllable in a word is supposed to

increase word duration by 178 ms in English and 213 ms in Mandarin if there is no tendency of equal word duration.

2.3.2.2.3 Syllable duration within word

To find out how syllables are compressed in words, I looked into syllable duration in terms of its position in word. In case the number of segments interacts with position, only CV syllables at word initial, medial and final positions were included in the analysis. In total 10841 English and 13571 Mandarin CV syllables were analysed. Figure 14A shows how syllable duration depends on stress and position in polysyllabic word in English (A) or in terms of percentage of monosyllabic word duration (B). Note that when calculating monosyllabic word duration, highly frequent words like “to” are excluded from analysis. The average monosyllabic CV word duration is 194 ms from my English data. Word-final syllables are longer than word-initial and word-medial syllables; word-initial syllables are slightly longer than word-medial syllables, and monosyllabic words behave similarly as word-final stressed syllables.

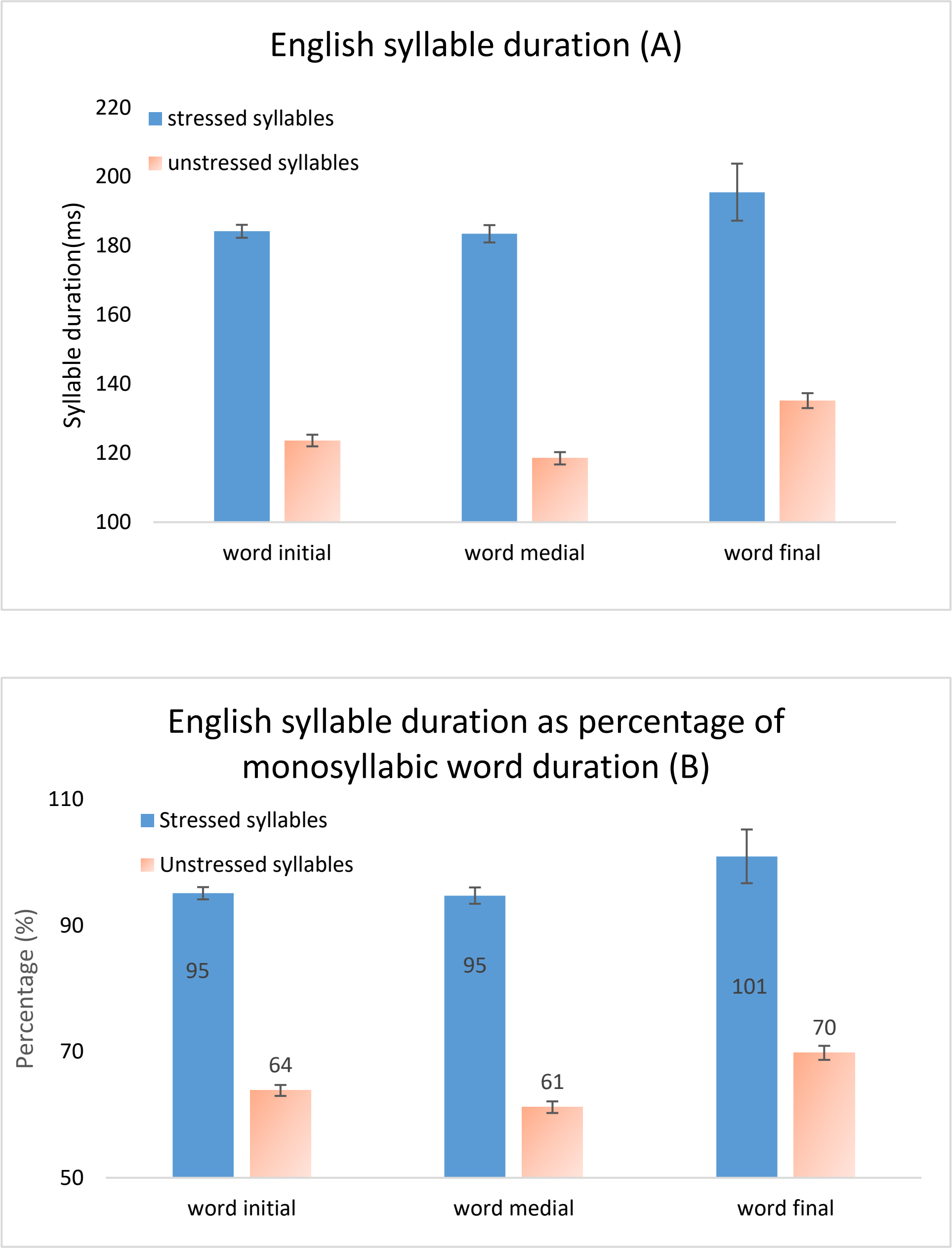


Figure 14. English syllable duration as a function of position in polysyllabic word, in (A) milliseconds, and (B) percentage of monosyllabic word duration. Only CV syllables at word initial, medial and final positions are included. Words that are at phrase-final or utterance final are excluded.

Mixed Model ANOVAs were performed, with stress and position (word initial, word medial, word final) as fixed factors, subject as random factor and syllable duration as dependent variable. The results showed a main effect of stress: F (1, 5.123) = 159.150, *p* < 0.001, partial *η*2 = .969, and a main effect of position, F (2, 11.621) = 10.005, *p* = 0.003, partial *η*2 = .623. The effect of subject is not significant, F (5, 4.612) = 0.274, *p* = 908, partial *η*2 = .229. There were interactions between stress and subjects, F (5, 14.440) = 4.771, *p* = 0.009, partial *η*2 = .623, and between stress, position and subject, F (10, 10805) = 3.591, *p* < 0.001, partial *η*2 = .03.

Bonferroni post-hoc analyses showed that word final syllables are significantly longer than word initial and word medial syllables. Although word initial syllables are slightly longer than medial syllables, they are not significantly different from each other. This is different from Nakatani et al.'s (1981) report that word-initial syllables were slightly but consistently longer than word-medial syllables. However, they did not support this observation with statistical analysis.

Figure 15 shows Mandarin syllable duration as a function of position in word either in milliseconds (A) or in percentage of monosyllabic word duration (B). The average monosyllabic CV word duration is 219 ms from my Mandarin data. Word-initial syllables are longer than word-medial and word-final syllables, and word-final syllables are longer than word-medial syllables. Mixed Model ANOVAs were performed on Mandarin data, with position (word initial, word medial, word final) as fixed factor, subject as random factor and syllable duration as dependent variable. The results showed a main effect of position, F (2, 21.101) = 160.133, *p* = 0.000, partial *η*2 = .938. The effect of subject is significant, F (9, 18.828) = 25.035, *p* =

0.000, partial *η*2 = .923. There was an interaction between position and subject, F (18, 15701) = 5.224, *p* = 0.000, partial *η*2 = .006. Bonferroni post-hoc analyses showed significant difference on each pairwise comparison between positions.

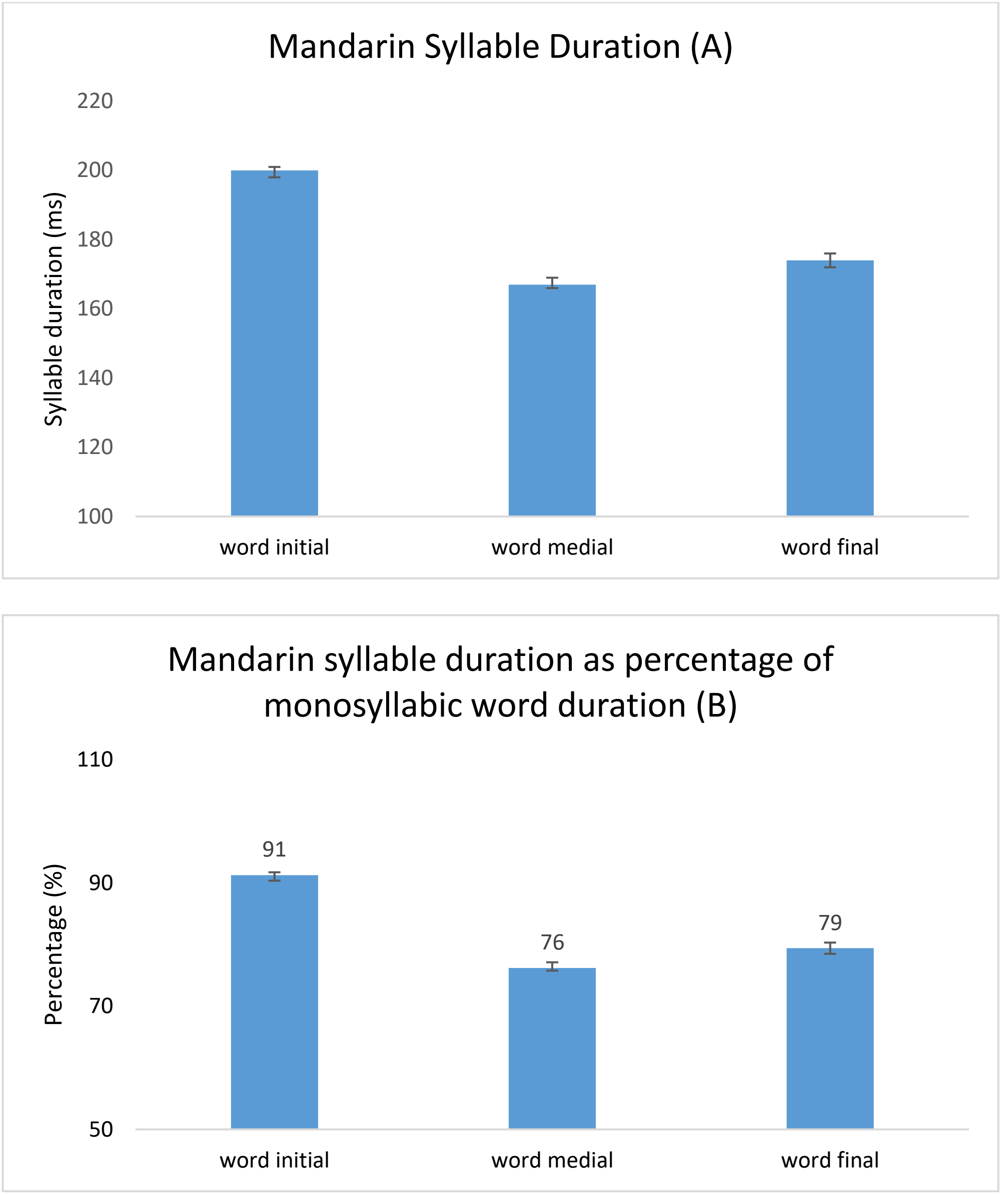


Figure 15. Mandarin syllable duration as a function of position in polysyllabic word, in (A) milliseconds, and (B) percentage of monosyllabic word duration. Only CV syllables at word initial, medial and final positions are included. Words that are at phrase-final or utterance-final were excluded.

Comparing Figures 14b and 15b, it can be observed that syllable duration is more compressible in Mandarin than in English. First, in English, word-final syllables are about equally long as monosyllabic words, whereas in Mandarin word-final syllables are much shorter than monosyllabic words. Second, in English, word-initial syllables are slightly but not significantly longer than word-medial syllables. In Mandarin, in contrast, word initial syllables are much longer than word medial syllables. The combined effects of word medial shortening and word final shortening, therefore, make Mandarin words much more compressible in duration than English words, for which both effects are absent.

What is also interesting is that word-final syllables in Mandarin are not only shorter than monosyllabic words, but also shorter than word initial syllables. Compared with mono syllabic words, word initial-syllables are 9% shorter, word-medial syllables are 24% shorter and word-final syllables are 21% shorter. This means that there is no word-final lengthening in Mandarin.

2.4 Discussion

This study has revisited the classical rhythm class hypothesis (Pike, 1945; Abercrombie, 1967) which posits that languages of the world are either stress-timed or syllable-timed (or mora-timed). It is true that experts have long accepted the fact that there is no direct evidence of isochrony in any language and they have long ceased to look for evidence of isochrony. However, despite failures to find evidence of equal timing at any of the alleged levels, the notion that languages are divided into timing-defined rhythm classes remains widespread and continues to drive rhythm-related research, and languages like English and Mandarin continue to be referred to as stress-timed or syllable-timed. In this study I have performed a more exhaustive search than before for evidence of timing-based rhythm through a corpus study aimed at identifying even the slightest tendency toward equal syllable or phrase duration in English and Mandarin, two languages that are, respectively, described as stress-timed and syllable-timed. This was done by controlling for linguistic functions that are known to significantly affect duration.

To remove doubts on the validity of findings based on controlled studies, I used two large non-experimental corpora, one in each language. This also enabled the examination of units larger than those investigated previously. I analysed the duration of segments, syllables, stress groups, words and phrases. The results not only showed confirmation of previous findings based on experimental data, but also additional information that, together, provided answers to the specific questions outlined in 2.1. To the question of whether English shows a tendency toward isochrony at any level, the answer is no. Once various other duration-affecting factors are controlled, inter-stress intervals were found to linearly vary their duration with the number of constituent syllables (Figure 11). Furthermore, phrase duration (Figures 12) and word duration (Figure 13) also varied linearly with their size. This indicates that, in English, syllables are not compressible (Figure 14) to show even a tendency toward equal inter-stress interval, equal word duration or equal phrase duration. This has removed the final trace of possibility that English is a stress-timed language based on isochrony. The reason for the incompressibility of English syllables has also become clear from my analysis. As shown in Figures 4-5, segments in English are also not compressible for the sake of equal syllable duration, whether measured in terms of intrinsic duration of segments (Figures 4-5) or number of constituent segments in a syllable (Figure 8).

To the question of whether Mandarin shows a tendency toward isochrony, the answer is not only yes, but also that it happens at both the syllable level and the phrase level. First, syllables showed a tendency toward equal duration, because their duration increased at a slower rate than the increase in the intrinsic segment duration (Figures 4-6) and in the number of constituent segments (Figure 8). This is consistent with the prediction of syllable timing for Mandarin. However, it is also found that syllables are compressible for the sake of equal word or phrase duration in Mandarin (Figures 12-13). This is not predicted by the syllable-timing classification of Mandarin, although it is not directly against it either.

It is important to note is that many of the present findings are not entirely new, as they have already been previously reported. The incompressibility of English segments is shown by van Santen and Shih (2000), though only in terms of linearity of syllable duration as a function of intrinsic duration of consonants and vowels. The incompressibility of segments in English is also implied in the linear relation between segment duration and syllable duration found by Crystal and House (1990) and O’Connor (1968). The high correlation between inter-stress interval duration and interval size can also be found in previous studies (Classe, 1939; Bolinger, 1965; Lea, 1974; Nakatani et al, 1981; O'Connor, 1965; Shen & Peterson, 1962). The lack of compressibility of syllables as a function of word size can be seen in Nakatani et al. (1981), which shows that there is no shortening of word final syllable relative to monosyllabic words (Figure 14). Although they observed that word-initial syllables were slightly longer than word-medial syllables, they did not provide statistical support. The compressibility of Mandarin segments in the direction of equal syllable duration can be seen in the data reported by van Santen and Shih (2000), although it was not the highlight of that paper. The compressibility of syllable duration in Mandarin for the sake of equal phrase duration can be seen in the findings of Xu and Wang (2009).

The present study, nevertheless, is the first to link up all these findings to reassess the tenability of the rhythm class hypothesis, and to cross-corroborate previous findings based on controlled speech data with current results based on non-controlled speech data. The evidence, as put together here, shows in unambiguous terms that the rhythm class hypothesis is simply untenable, because English, as an exemplary prototypical stress-timed language, and one of the original languages based on which the hypothesis was proposed, shows no trace of possibility for isochrony of stress intervals to happen, whereas Mandarin, an alleged and “confirmed” (based on rhythm metrics) syllable timed language, does show a small tendency toward isochrony at the word level and phrase level. The present results have also revealed the source of this inverse result: the lack of compressibility of segments in English versus compressibility of segments in Mandarin. The compressibility of Mandarin segments makes it possible for syllables to be shortened or lengthened toward not only equal syllable duration, but also equal word duration and equal phrase duration. The lack of segmental compressibility in English when major functional linguistic factors such as lexical stress and boundary strength are taken into consideration, in contrast, makes it impossible to adjust syllable duration to achieve even a tendency toward isochrony of stress intervals. Based on these results, I agree with doubts from previous studies (Arvaniti and Rodriquez, 2013; Nolan and Jeon, 2014; White et al, 2012) and I would like to suggest that it is time to abandon the rhythm class hypothesis.

2.5 Summary

It has been long held that languages of the world are either stress-timed, syllable-timed or mora-timed. By definition (Abercrombie, 1967), in a stress-timed language, inter-stress intervals are constant, hence, isochronous, while in a syllable-timed or mora-timed language, successive syllables or morae are equal in duration. Empirical research so far, however, has failed to find evidence of such isochrony in any language, yet the hypothesis is now sustained by perception accounts or rhythm metrics that do not measure isochrony (Dellwo, 2006; Dellwo & Wagner, 2003; Grabe & Low, 2002; Ramus et al., 1999). Here I have reexamined the rhythm class hypothesis by looking for evidence of at least a tendency toward isochrony, through a comparison of English, a prototypical stress-timed language, and Mandarin, an alleged (Lin & Wang, 2007) and “confirmed” (Grabe & Low, 2002; Lin & Wang, 2007; Mok & Dellwo, 2008) syllable-timed language. I examined the relationship between segment and syllable duration and the relationship of syllable and phrase duration in one large spoken corpus from each language. The results show that in English, segments are incompressible to allow equal syllable duration, and syllables are incompressible to enable equal phrase duration. There is therefore no tendency toward stress timing in English. In contrast, Mandarin shows a small tendency toward equal syllable duration, equal word duration and equal phrase duration. These findings are exactly the opposite of what would be predicted by the rhythm class hypothesis. I therefore argue that the hypothesis is not just weak, as it has so far been criticized, but simply untenable, and the so-called rhythm class divisions should no longer be held as a basic fact of human language.

3 Mandarin and English use different temporal means to mark major prosodic boundaries

3.1 Background

Boundary is a function that allows speakers to divide continuous speech into smaller chunks for easier auditory comprehension. Apart from pre-boundary lengthening and silent pause mentioned in the literature review, there are other cues that can also be used to mark boundaries. It has been reported that the duration of consonants is longer in word-initial position than that in word-medial position (Oller, 1973; Cooper, 1991; Fougeron and Keating, 1997; White and Turk, 2010). All languages may have boundary marking lengthening (Beckman, 1992, Fletcher, 2010). It is also found that there is pitch reset at certain boundaries (Ladd, 1988; Shen, 1990; Shih, 2000; Swerts and Ostendorf, 1997). Among the cues reported for boundary marking, pre-boundary lengthening and silent pause are the most important ones (Lehiste, 1972; Xu, 2009). However, it is unclear how exactly these two types of cues are distributed across boundaries of different strengths.

Boundary strength is often represented by break index in the ToBI annotation system (Beckman & Ayers, 1997). In this system, break level 0 refers to syllable boundaries within word; level 1 refers to normal word boundary (or most phrase-medial word boundaries); break level 2 refers to a lower-level perceived grouping of words without an intermediate B3 or a

B4 intonational boundary marker or the break with next word is weaker than expected although the pitch pattern clearly suggests an intermediate B3 or a B4 intonation phrase boundary; and break levels 3 and 4 are largely defined by intonational phrasing, referring to intermediate phrase and full intonation phrase, respectively (Beckman & Ayers, 1997).

For Mandarin, a slightly different break index system, namely, C-ToBI, is widely used (Li, 2002). In C-ToBI, break level 0 indicates the minimum break between syllables, usually within a prosodic word; level 1 indicates prosodic word boundary; level 2 refers to minor prosodic phrase boundary; level 3 refers to major prosodic phrase boundary; and level 4 refers to prosodic group boundary (Li, 2002). A major difference between C-ToBI and ToBI is that C-ToBI labelers do not need to use pitch accents and boundary tones to mark break indices in Mandarin.

There has been some research on the relationship between boundary strength and pre-boundary lengthening, particularly for high-level boundaries in English. For example, by using reiterant-speech (syllable /no/) versions of sentences with different phrasings read by three American English speakers, Fougeron and Keating (1997) found that domain-final vowel lengthening is weakly cumulative. Two levels of boundaries can be distinguished by duration for two speakers (1 and 2), intonational phrase-final and intermediate phrase-final being longer than word-final and syllable-final /o/'s. All four levels can be distinguished for speaker 3. Wightman, Shattuck-Hufnagel, Ostendorf and Price (1992) looked at a speech corpus that has 35 pairs of sentences that are phonetically similar but have different meanings. Those sentences were read by four professional newsreaders. Their study showed significantly different amounts of pre-boundary lengthening in American English between all four levels of boundary strength: prosodic word, a group of words within a larger unit, intermediate phrase, and intonational phrase.

Mandarin showed no significant difference in pre-boundary lengthening between minor prosodic phrase and major prosodic phrase boundaries. Yang (1997) examined the prosodic cues of syntactic boundaries in 48 designed sentences. The average length of her sentences is 10 words. In her research, break level 1 indicates syllable boundary, break level 2 indicates word boundary; break level 3 indicates word group boundary; break level 4 indicates phrase boundary; break level 5 refers to clause boundary and level 6 refers to sentence boundary. She found that the duration of the syllable increases at first before word group and phrase boundaries, and then decreases before clause and sentence boundaries. Subsequent studies (Cao, 2005: Li, 1998; Yang & Wang, 2002). corroborated some of her findings regarding pre-boundary lengthening. According to Li (1998) and Yang and Wang (2002), there is no significant difference in pre-boundary lengthening between minor prosodic phrase and major prosodic phrase boundaries. Cao (2005) found that prosodic phrase final syllable is significantly longer than average syllable. Whereas the sentence-final syllable has a duration that is comparable to or less than average syllable duration. Although these researchers use different break annotation systems, it appears that pre-boundary duration does not increase linearly with break level in Mandarin. This raises the question of whether there is a difference between English and Mandarin in terms of pre-boundary lengthening.

Silent pause and boundary strength are also related. Mandarin normally has no silent pause after prosodic words. However, silence duration grows with boundary strength (word (Qian et al., 2001; Xiong, 2003; Yang & Wang, 2002). This seems to suggest a trade-off relation between pre-boundary lengthening and silent pause for larger boundaries in Mandarin. In English, 23% of the "intonation phrase" boundaries contained an unfilled pause, and 67% of the "groups of intonation phrases" have an unfilled pause (Wightman et al., 1992). Lehiste (1979) and Scott (1982) implied that lengthening and pausing may balance each other.

Some studies have proposed to combine pre-boundary duration and pause durations. Pre-boundary duration and pause duration affect the temporal distance between the onsets of the pre-boundary constituent and the post-boundary constituent (Xu & Wang, 2009). The temporal distance is then suggested by Xu and Wang (2009) to indicate relational distance between neighbouring constituents. Yang (1997) found that in Mandarin, the sum of pre-boundary syllable duration and following pause duration increases linearly with the level of boundaries. However, she only used 48 sentences. And their sentence-final pause is problematic. Because her sentences are isolated, measuring sentence-final pause is difficult. More data in the form of paragraph is needed to further confirm the results in

Mandarin. It is also worth looking into English to see if temporal distance can distinguish different break levels.

The present study is a preliminary corpus analysis aimed at examining the relationship between boundary strength and pre-boundary lengthening and silent pause in English and Mandarin. I particularly want to know whether there is a difference between English and Mandarin in terms of pre-boundary lengthening and whether temporal distance can distinguish break levels in both languages.

3.2 Methods

3.2.1 English Corpus

For English, just like in Chapter 2, the Boston University Radio News Corpus was used (Ostendorf et al., 1995). The description of this corpus can be seen in Section 2.2.1. Figure 16 is an example of break index labelling from the corpus.

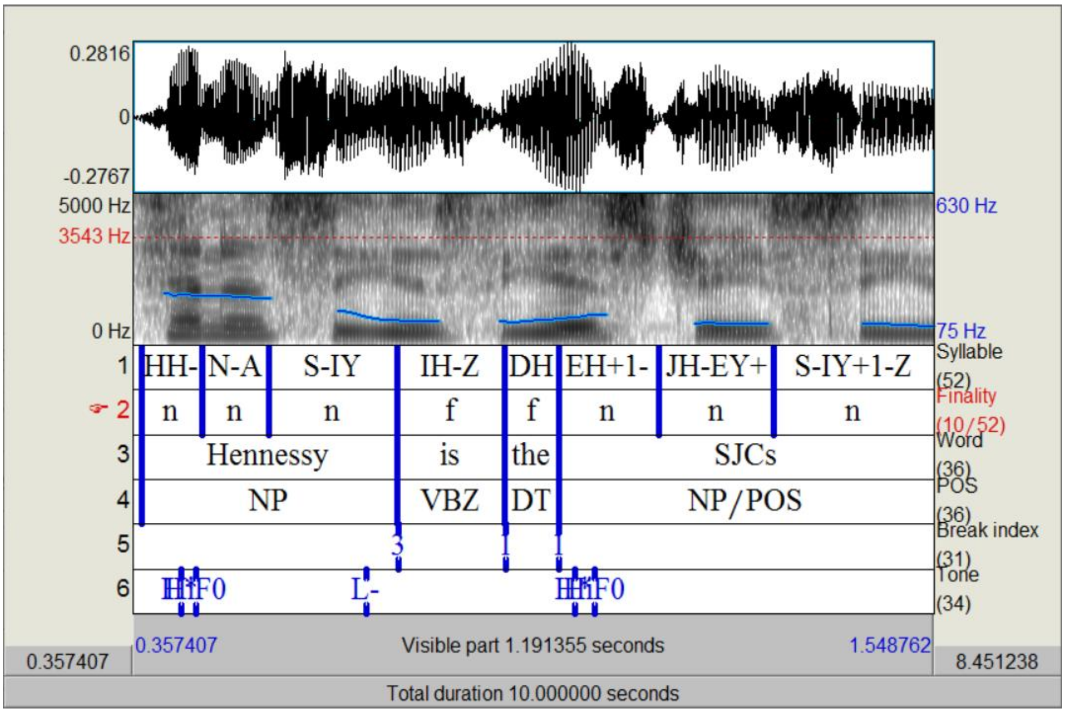


Figure 16. Illustration of break index labelling in English.

Not every sentence in the corpus is annotated. Only sentences with complete segment, syllable, and prosodic information were used. In the end, 369 sentences which have both prosodic labels and syllable information were analysed in this study. Table 4 shows number of tokens before different break indices for each speaker in the English corpus.

Table 4. Number of tokens before different break indices for each speaker in English corpus.

Speakers	B1	B2	B3	B4
f1	2382	198	345	698
f2	6683	1585	1102	2757
f3	1692	78	276	422
m1	1808	262	278	389
m2	1703	135	256	537
m3	1358	26	142	304

3.2.2 Mandarin Corpus

The Mandarin data were from the Annotated Speech Corpus of Mandarin Discourse (ASCCD), which is described in Section 2.2.2. Break indices had been previously labelled using C-ToBI (Li, et al., 2000). Syllables with a neutral tone were excluded from the analysis. Figure 17 is an example of break labelling in this Mandarin corpus. Some data from speaker m2 could not be shown in Excel properly when extracting pre-boundary syllable duration, so his data was not included in the following analysis. Table 5 shows the number of tokens before different break indices for the other 9 speakers in the Mandarin corpus.

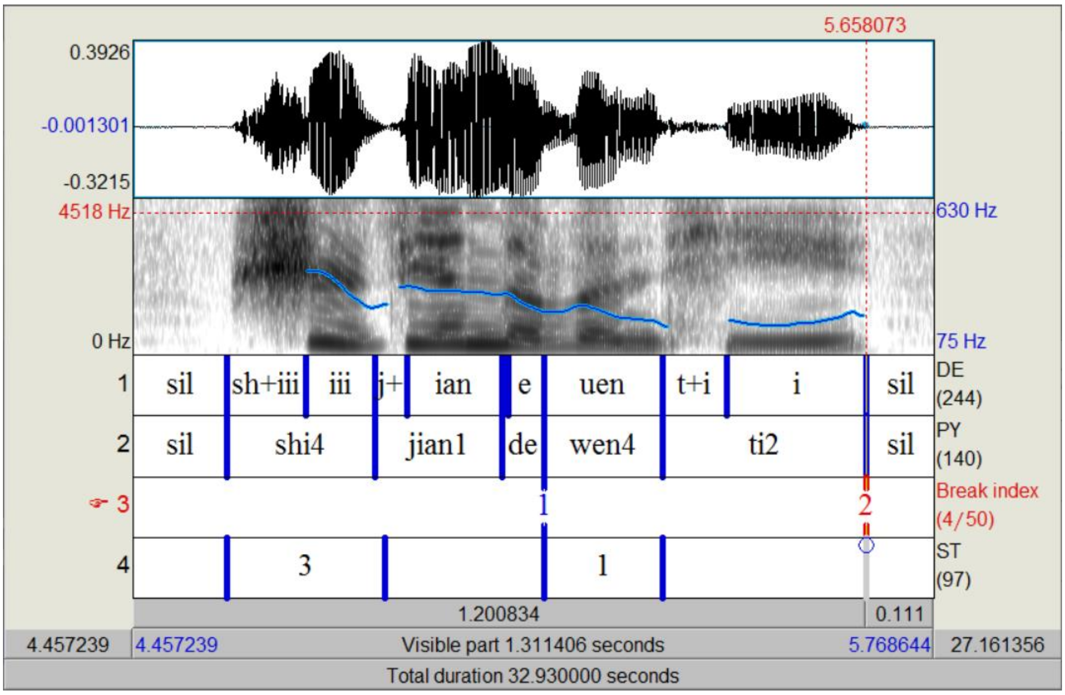


Figure 17. Illustration of break index labelling in Mandarin. "Problems in the world…"

Table 5. Number of tokens before different break indices for each speaker in Mandarin corpus.

Speakers	B1	B2	B3	B4
f1	1395	588	504	325
f2	1140	598	360	269
f3	1011	533	462	221
f4	1093	559	485	269
f5	436	893	430	157
m1	1146	516	675	380
m3	1512	682	663	338
m4	955	607	448	242
m5	1142	692	850	241

3.2.3 Measurement

The measurements made were pre-boundary syllable duration, silence duration, and their sum as temporal distance.

3.3 Results

3.3.1 Mandarin results

The mean duration patterns are shown in Figures 18-20. Figure 18 shows that pre-boundary syllable duration ceases to lengthen beyond break level 2. In contrast, as shown in Figure 19, temporal distance, which combines silent pause and pre-boundary duration, continues to increase beyond break level 2 for both monosyllabic and polysyllabic words.

A two-way repeated measures ANOVA was conducted, with the number of syllables (1 or more) in pre-boundary words and break index (1, 2, 3 and 4) as within-subjects factors, pre-boundary syllable duration as the dependent variable. All reported ANOVAs are by subjects in this chapter. The results showed a main effect of the number of syllables: F (1, 8) = 81.059, p < 0.001, partial η2 = .910, and a main effect of break index, F (3, 24) = 37.714, p < 0.001, partial η2 = .825. There was no interaction between the number of syllables and break index.

Bonferroni post-hoc analyses revealed that pre-boundary syllable before break 1 (M = 0.197, SD = 0.005) was significantly shorter than that before other breaks (break 2, M = 0.256, SD = 0.11, break 3, M = 0.264, SD = 0.009, break 4, M = 0.252, SD = 0.008). However, the other break levels do not differ from each other on pre-boundary syllable duration.

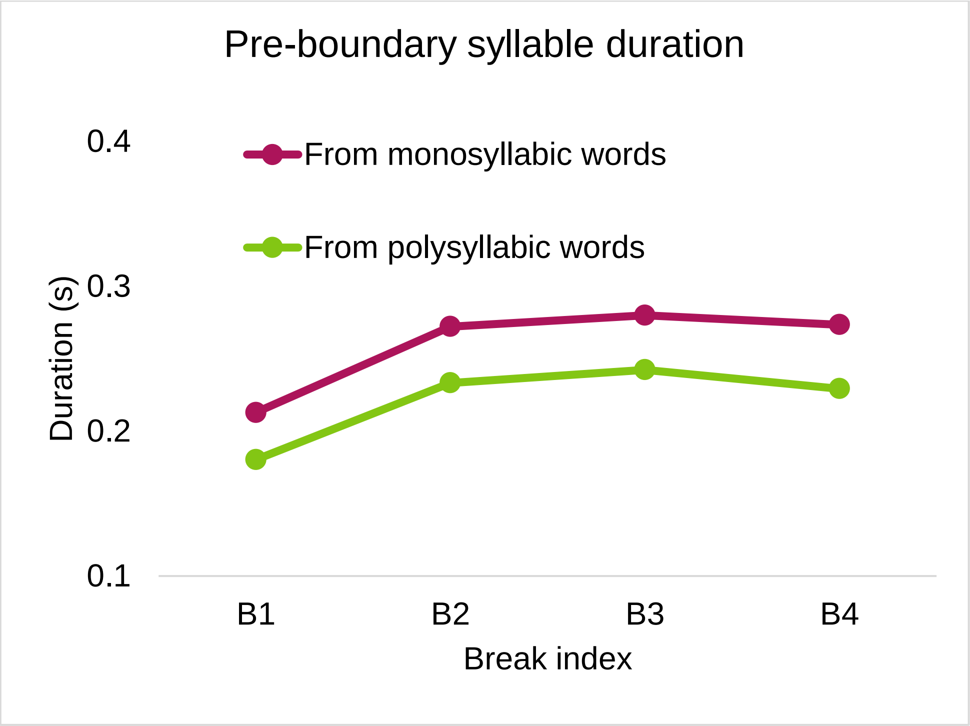


Figure 18. Pre-boundary syllable duration as a function of break index in Mandarin.

A two-way repeated measures ANOVA was conducted with the number of syllables in pre-boundary words (1 or more) and break index as within-subjects factors, pre-boundary syllable duration as the dependent variable. The results showed a main effect of the number of syllables, F (1, 8) = 60.367, *p* < 0.001, partial η2 = .883, and a main effect of break index, F (1.151, 9.206) = 71.612, *p* < 0.001, partial η2 = .900.

Bonferroni post-hoc analyses showed significant difference in each pairwise comparison between temporal distance at break 1 (M = 0.202 s, SD = 0.006 s), break 2 (M = 0.313 s, SD = 0.017 s), break 3 (M = 0.686 s, SD = 0.039 s) and break 4 (M = 0.894 s, SD = 0.075 s), *p* < 0.05. This means that the larger boundary shows significantly longer temporal distance than the smaller boundary in each pair of comparison.

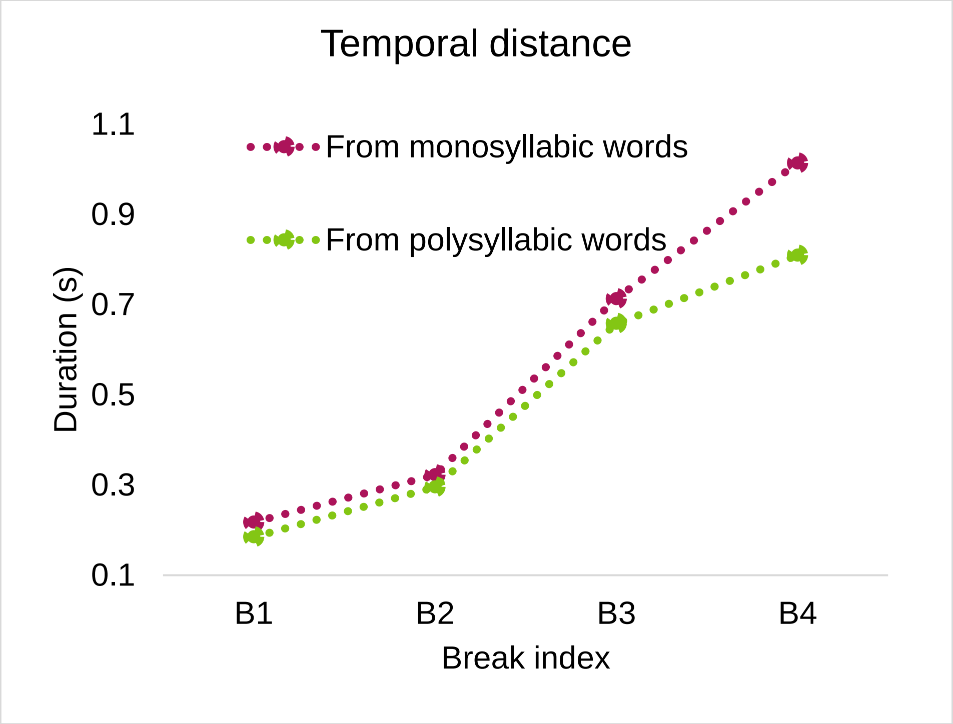


Figure 19. Temporal distance as a function of break index in Mandarin.

There is an interaction between number of syllables and break index, F (1.197, 9.578) = 10.456, $p < 0.001$, partial η2 = .567. A follow-up Paired-Samples t-Test showed that all paired samples are significantly different, $p < 0.05$. The effect of break index was more pronounced in syllables from monosyllabic words than polysyllabic words as break index increased.

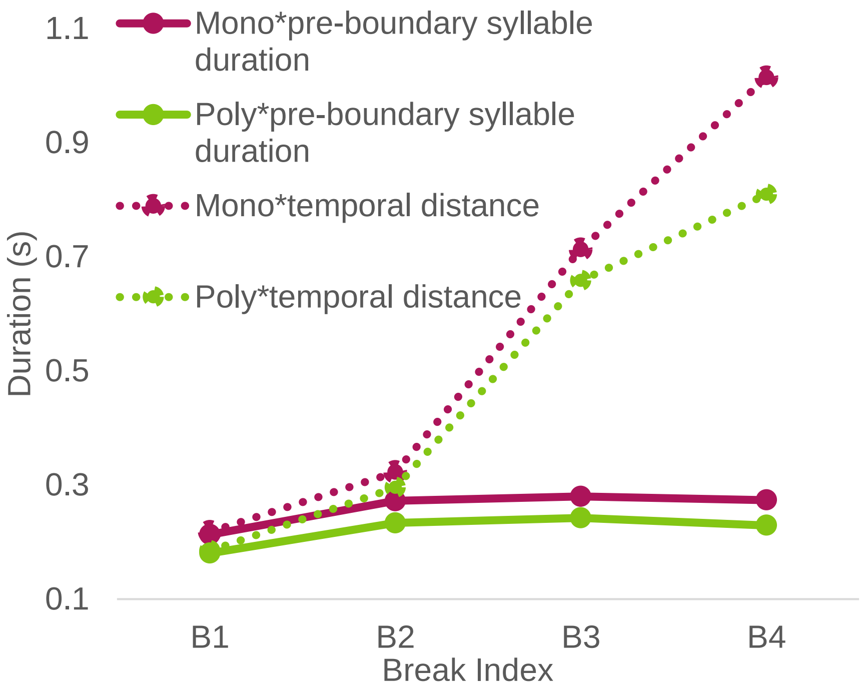


Figure 20. Pre-boundary syllable duration and temporal distance over break index in Mandarin.

3.3.2 English results

In this section, I report results from monosyllabic words and polysyllabic words separately. This is because F3 and M3 have less monosyllabic tokens before B2 and thus intrinsic of duration of segments and number of segments could override other effects. To reduce the impact of data sparsity, I only present the results of monosyllables without separating stressed and unstressed conditions. Function words are included in the following analysis.

3.3.2.1 Monosyllabic Words

Monosyllabic words including function words before each break index (1, 2, 3 and 4) are analysed in this section. Figure 21 shows that pre-boundary syllable duration increases gradually over break levels. It also shows that temporal distance has a similar trend and is largely overlapped with pre-boundary syllable duration except for break level 4.

Repeated-measures ANOVAs on pre-boundary syllable duration and temporal distance were conducted, with break level as a fixed factor and subjects as a replication factor. As is shown in Table 6, the results showed significant effects of break index on both pre-boundary syllable duration and temporal distance. Bonferroni post-hoc analyses revealed that each pairwise comparison was significant, $p < 0.05$. This suggests that the larger boundary shows significantly longer pre-boundary syllable duration and temporal distance than the smaller boundary in each pair of comparison, as is shown in Table 6.

Table 6. Results of repeated measures ANOVAs on the effect of break index on pre-boundary syllable duration and temporal distance. The unit of measurement is second.

Pre-boundary syllable duration	Temporal distance
F (3, 15) = 72.937, p < 0.001.	F (1.108, 5.540) = 38.903, p < 0.01.
B1 (0.160),	B1 (0.162),
B2 (0.223),	B2 (0.232),
B3 (0.297),	B3 (0.301),
B4 (0.350)	B4 (0.444)

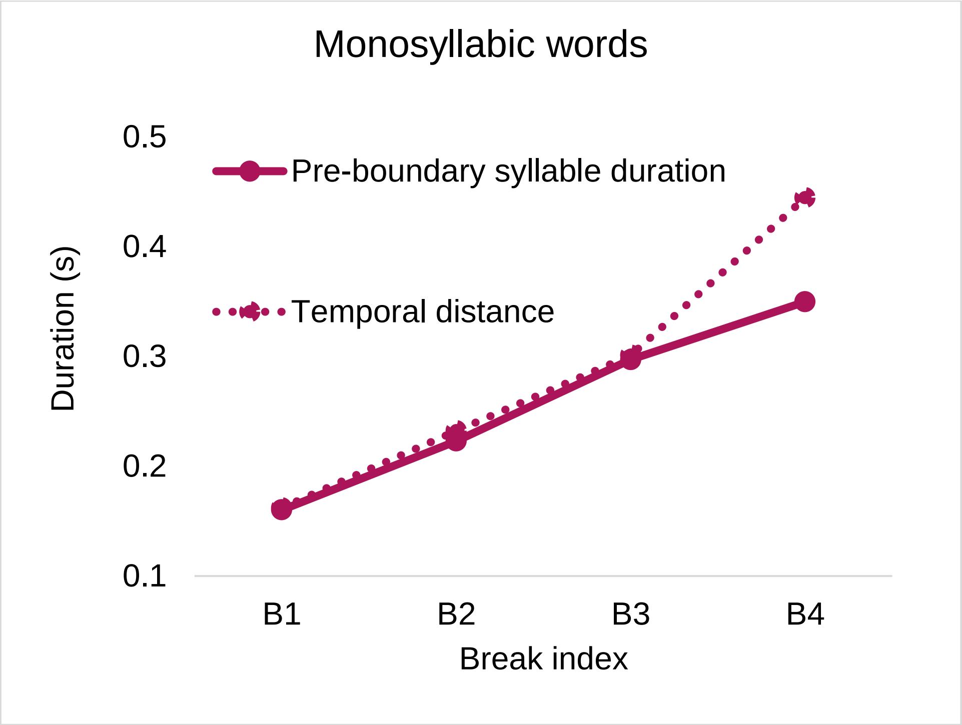


Figure 21. Pre-boundary syllable duration and temporal distance over break index after monosyllabic words in English.

3.3.2.2 Polysyllabic Words

Syllables from polysyllabic words including function words before each break index (1, 2, 3 and 4) are analysed in this section. Figure 22 shows that pre-boundary stressed and unstressed syllable duration increases gradually over break index. Also, temporal distance has a similar trend and is largely overlapped with pre-boundary syllable duration except for break level 4.

Repeated-measures ANOVAs on pre-boundary syllable duration and temporal distance were conducted with stress (stressed and unstressed) and break index as fixed factors and subjects as replication factor. As is shown in Table 7, the results indicated a main effect of stress and a main effect of break index on both pre-boundary syllable duration and temporal distance. There was no interaction between the two factors. Bonferroni post-hoc analyses showed that each pairwise difference was significant, p < 0.05. This means that the larger boundary shows significantly longer pre-boundary syllable duration and temporal distance than the smaller boundary in each pair of comparison, as is shown in Table 7.

Table 7. Results of repeated measures ANOVAs on the effect of break index and stress on pre-boundary syllable duration and temporal distance in English. The unit of measurement is second.

	Pre-boundary syllable duration	Temporal distance
Break index	F (3, 15) =90.651, p < 0.001. B1 (0.195), B2 (0.221), B3 (0.253), B4 (0.299)	F (1.117, 5.587) = 58.528, p < 0.001. B1(0.199), B2(0.223), B3(0.258), B4(0.403)
Stress	F (1, 5) = 303.664, p < 0.001	F (1, 5) =1309.778, p < 0.001

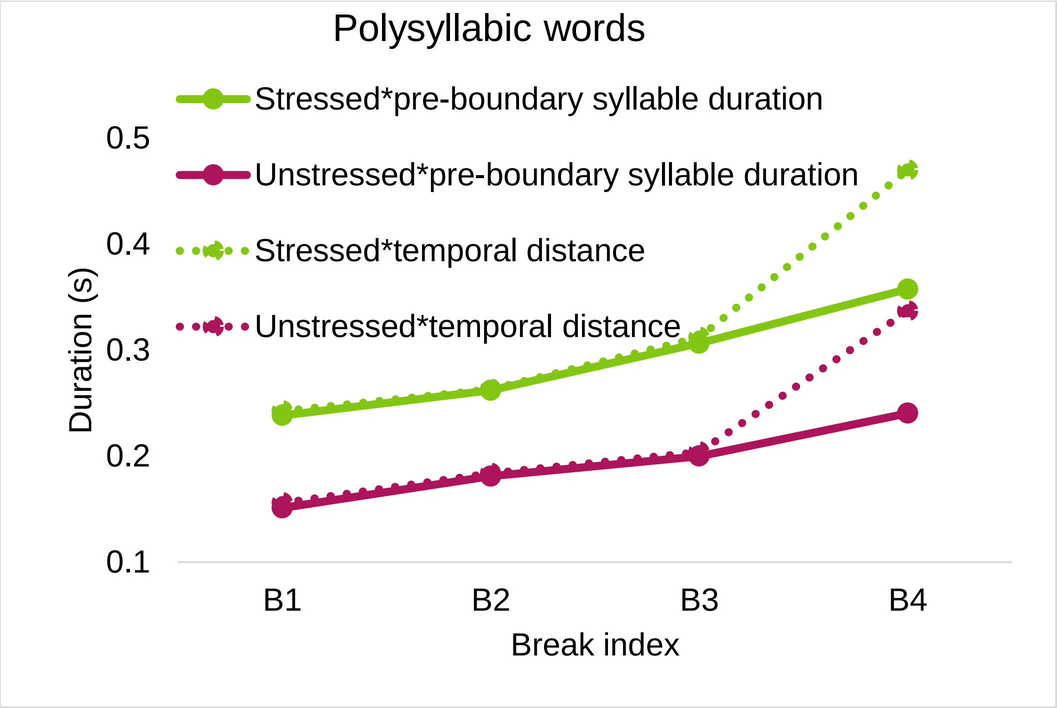


Figure 22. Pre-boundary syllable duration and temporal distance over break index after polysyllabic words in English.

3.4 Summary

To highlight the main finding of the study, Figure 23 plots pre-boundary syllable duration in both Mandarin and English. As can be seen, in English, pre-boundary syllable duration increases continuously with break index, whereas in Mandarin, the duration increase stops beyond break 2. This is consistent with previous reports for Mandarin (Li, 1998; Yang, 1997; Yang & Wang, 2002) and English (Wightman et al., 1992), respectively. In the current study, the difference between the two languages is directly demonstrated: the duration of silent pause significantly increases beyond break level 2 in Mandarin, as if to compensate for the lack of continuous syllable lengthening, while the increase in English is less significant.

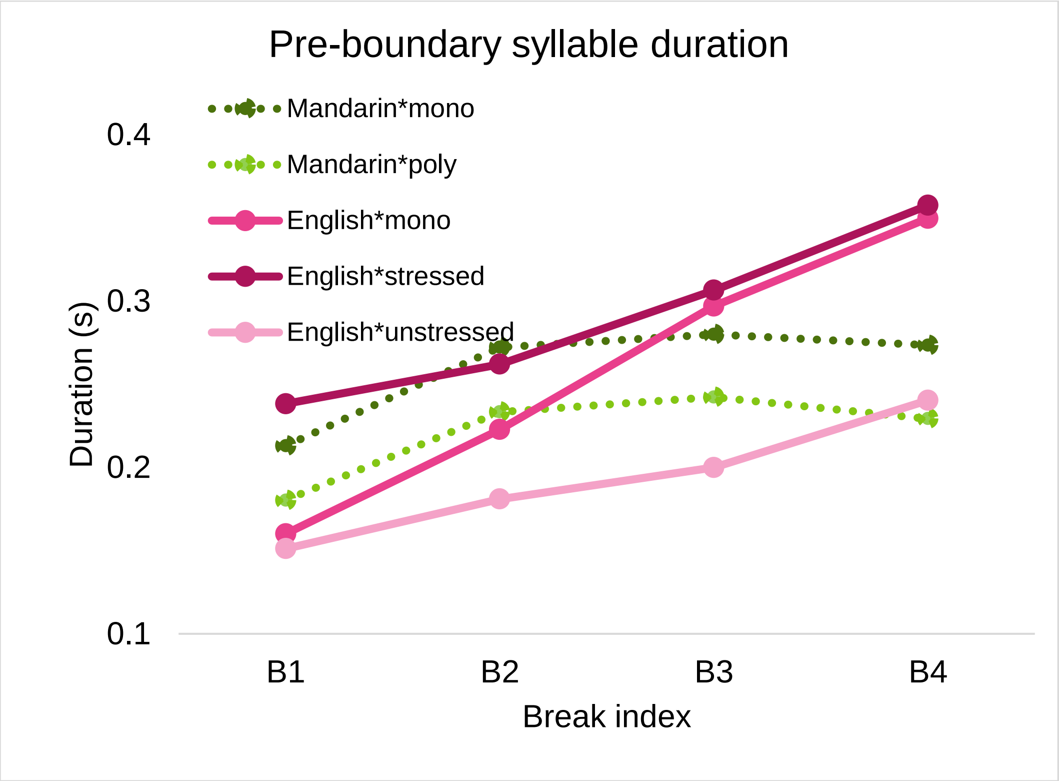


Figure 23. Pre-boundary syllable duration in English and Mandarin as a function of break index.

4 Effects of Part of Speech: Primitive or derived from word frequency?

4.1 Background

As mentioned in 1.5, POS is an important input in both speech recognition and speech synthesis modelling. However, it is unknown why it is so important. It is possible that POS may influence duration and F0. The POS effect has been studied in some experiments. Sorensen, Cooper and Paccia (1978) found that nouns are longer than verbs in typical sentences when noun-verb homophones in sentences are matched for phonetic environment and stress pattern. They claimed that this difference is caused by the fact that nouns are from a larger grammar category than verbs, so the information load carried by a given noun is larger than that carried by a verb, under the assumption that duration is a positive correlate of information load which was indicated by frequency of occurrence (Coker et al., 1973). However, the study did not take grouping and frequency into account. The frequencies of nouns and verbs are different in each pair. Also, most of the nouns are phrase-final, and verbs are phrase-initial. The POS effect they found could be confounded by the effects of frequency and final lengthening.

There have also been comparisons of function words and content words. Most function words are highly frequent, while most content words are rare (Meunier & Espesser, 2011). Johnson (2004) did a segmental deviation analysis, which is a bit more sensitive than syllable deletion analysis, in a corpus of spontaneous conversations between American English speakers. He found asymptotic deletion and deviation rates of 20% deletion and 40% deviation for function words and 10% deletion and 25% deviation for content words. This indicates that function words are deleted more often than content words. Meunier and Espesser (2011) investigated the influence of word category in a corpus of French conversational speech and found a clear effect of word category on vowel duration: content words showed longer vowel duration, while function words showed shorter duration. However, the function words in French appear more often in unstressed position, so the word category effect was confounded with position and stress.

The aforementioned research did not consider frequency of occurrence, but POS is related to it (Meunier & Espesser, 2011). It is possible that the POS effect is a result of word frequency. Frequency of occurrence is also important in information theory, as it is directly related to information load (Shannon, 1948). If there is a frequency effect on duration, this effect can find support from information theory.

The frequency effect on word length (Zipf, 1935) is widely known. Piantadosi, Tily and Gibson (2011) proposed some modification to Zipf's law. The rationale behind this is that they believe Zipf's law would work if words occur randomly from a stable distribution, but in reality, word probabilities vary a lot based on context due to the nonstationary nature of natural language. Therefore, an improved code for meanings can be developed by taking into consideration the statistical relationships between words. Then the authors used joint probability and N-gram to calculate the information content, which they refer to as average information content. They found that average information content is a better predictor of word length than frequency. However, although they did partial correlations in statistics, joint probability and N-gram are related to word frequency by definition, so that frequency may still have confounded the information content effect. This means that word frequency still has an impact on word length.

Both Zipf (1935) and Piantadosi et al. (2011) are mainly focused on the number of syllables in a word. In real speech, however, syllable deletion or segment deviation may also occur as a phonetic process (Johnson, 2004). Sometimes the realized segments cannot match the citation form. This has led to my hypothesis that the sensitivity of code length in speech occurs not only at the lexical level, but also at the phonetic level. In fact, it is possible that code length variation starts from phonetic duration, which in turn leads to variation in phonetic reduction. And phonetic reduction may eventually lead to a reduction of word length when entire syllables are no longer sustained.

Some studies have reported the POS effect on F0. Shih (2000) discovered that verbs have lower F0 values than surrounding polysyllabic nouns in Mandarin data, resulting in shallow F0 valleys. She believes that verbs are metrically weaker than nouns in Mandarin. Bulut et al. (2007) found that in English, the POS tags in their first half of sentences have higher F0 median values than the POS tags in their second half of sentences. Both studies did not consider position as a potential confounding factor in their data. It is worthwhile to investigate the POS effect while controlling for the position factor. Additionally, there has also been evidence of frequency effect on F0. Zhao and Jurafsky (2009) proposed that the tonal production of a speaker exists in a dynamic tone space, with each production having a unique tonal dispersion from the tonal centre. They demonstrated that low-frequency words had a generally more dispersed tone space than high-frequency words.

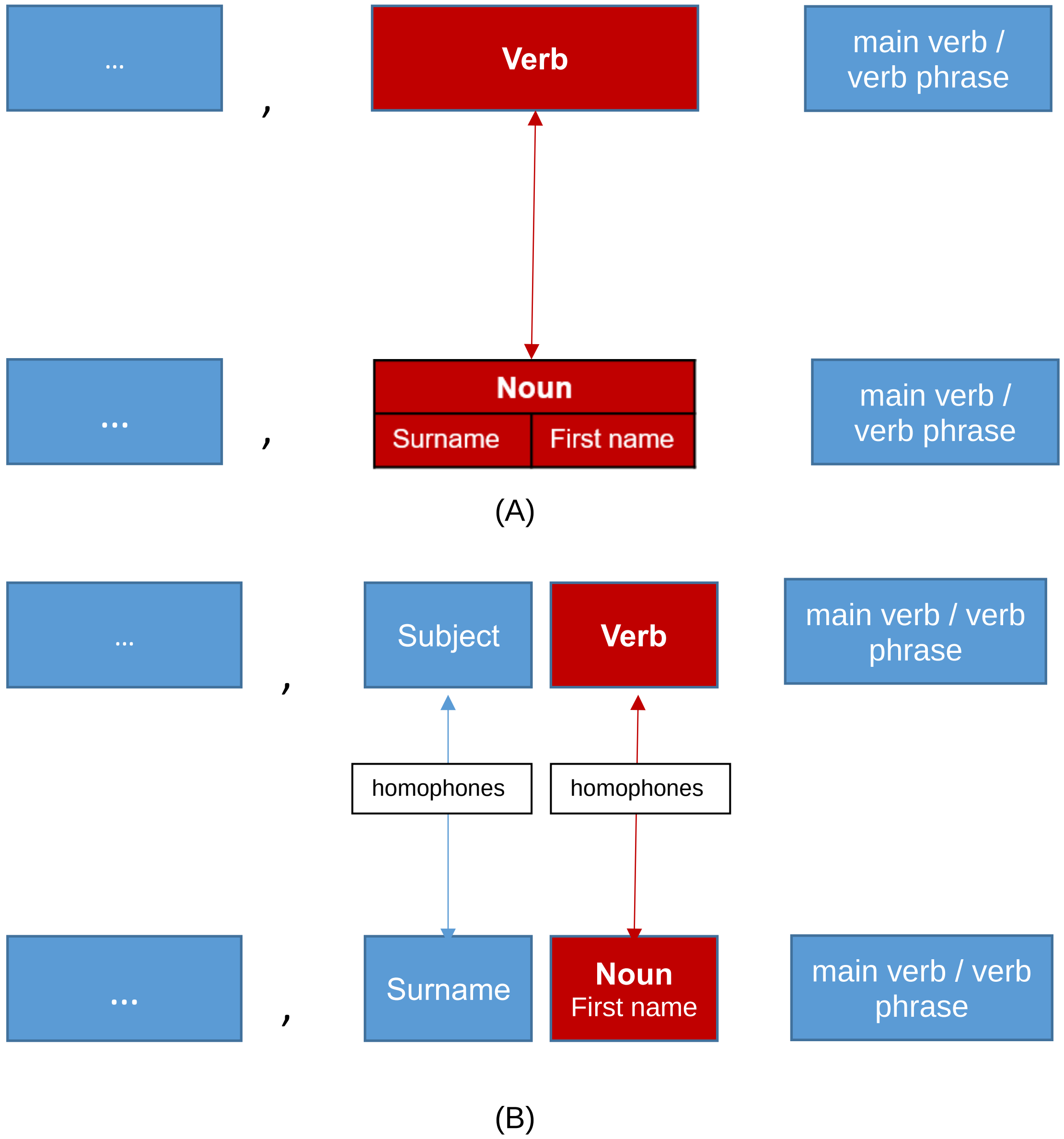
The purpose of this study is to establish a clearer link between the effects of POS and word frequency on duration and F0 than has been proposed before. My hypothesis is that the POS effect on duration and F0 is actually derived from the effect of word frequency. This is done by comparing pairs of nouns and verbs in Mandarin that are phonetically identical.

4.2 Method

4.2.1 Stimuli

In this experiment, 44 pairs of verb and noun (12 of them are monosyllabic and 32 of them are disyllabic) are embedded in designed sentences. The sentence design made use of the serial verb construction (Yin, 2007) in Mandarin. This means that there are a series of verbs or verb phrases in a sentence without the use of an intervening conjunction. Figure 24 shows the sentence design and Table 8 illustrates some examples of the sentences that were made with this design. The nouns are made-up names. To control the boundary effect, the verb and noun in each pair are at the same position in each pair of sentences. Each pair of sentences shares the same sentence length and grouping pattern. Homophones of a verb and a noun were used to control the effect of syllable structure. There was no prosodic focus on the target words. The complete list of sentences can be found in Appendix 1 and the complete list of words are in Appendix 2.

The reason to use made-up names is that they make it possible to compare a noun and a verb in two grammatically correct and meaningful sentences that share the same sentence length and grouping pattern. The syllables in the latter half of the sentences are as similar as possible, if not identical.



(A)

(B)

Figure 24. Sentence design. (A) illustrates cases in which a homophone of the initial syllable of the verb is a surname. The sentence structure in (A) is … + verb + main verb / verb phrase for verbs, and … + main verb / verb phrase for nouns. If a homophone of the initial syllable of the verb is not a surname, a surname is inserted before the noun and a subject before the verb, as shown in (B). In this scenario, the inserted subject and the surname are homophones or nearly homophones. The sentence structure in (B) is … + subject + verb + main verb / verb phrase

for verbs, and … + subject + main verb / verb phrase for nouns. One may be concerned about the grouping pattern in (B). The verb is always attached to the preceding subject in Mandarin. The noun here is a made-up first name which is attached to the preceding monosyllabic surname. So the target words and preceding words are closely related.

Table 8. Illustrations of sentences. “ai” and “kaishi” are the target words.

Sentences
这条街的邻居都知道，你爱（v.）画画。 Zhetiaojiede linju dou zhidao, ni***ai*** huahua. All the neighbours in this neighbourhood know that you ***like*** drawing.
为了表达自己的想法，李艾（n.）画画。 Weile biaoda zijide xiangfa, Li***ai*** huahua. Li***ai*** paints to express her idea.
为了提高身体素质，你开始（v.）学习游泳。 Weile tigao shenti suzhi, ni ***kaishi*** xuexi youyong. To improve health condition, you ***start to*** learn swimming.
为了提高身体素质，李开使（n.）学习游泳。 Weile tigao shenti suzhi, Li ***kaishi*** xuexi youyong. To improve health condition, Li ***kaishi*** learns swimming.

The nouns were designed in the following way. Take 爱 ai (tone 4, verb, love) for example. To minimize the effect of segment identity, number of segments and tone, it is necessary to have a homophone as the counterpart of 爱 ai (tone 4, verb, love). I chose 艾 ai (tone 4, noun, name) which is a name that is used by some people in China. Then I need to put both 爱 ai (tone 4, verb, love) and 艾 ai (tone 4, noun, name) in sentences. To minimise the effect of neighbouring syllables, 画画 huahua (draw; draw pictures; paint a picture;) is used in both sentences. 你(ni tone 3, you) is a pronoun. Although it is not the target word, to minimise noise as much as possible, it is ideal to find a surname that is a homophone of 你 (ni tone 3, you). However, there is no such surname in Mandarin, so I used 李 (li tone 3) which is a common surname. Boundaries are controlled too. Here the verb (爱, love) is attached to the pronoun (你, ni, you), and the first name (艾, ai) is attached to the last name (李, li). How sentences for 爱 ai and 艾 ai are constructed based on the sentence design is shown in Figure 25.

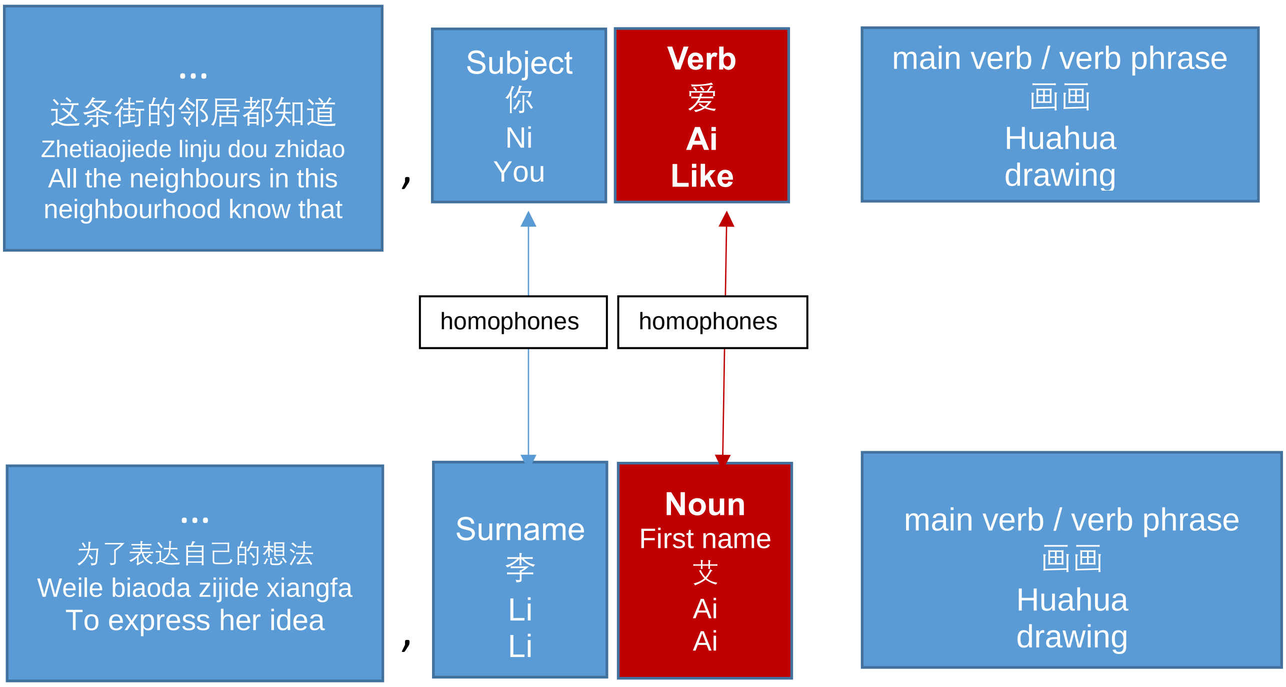


Figure 25. How sentences for 爱 ai and 艾 ai are constructed based on the sentence design in this study.

12 pairs of verb and noun are monosyllabic and 32 pairs are disyllabic. If the noun counterpart is extremely rare and unnatural to be a name, some changes about tones or segments had to be made to find more appropriate nouns (eg. 怕 pa4/ 坡 po1, 乐 le4 意 yi4/ 陆 lu4 毅 yi4). When choosing these nouns, segments that share similarities with those from verbs were considered to control the effect of segment identity.

4.2.2 Participants

Five female and four male native Mandarin speakers were recorded. They were all college students studying in London at the time of recording, and they were born and raised in Beijing. None of them reported any history of speech or hearing impairment. The recordings were conducted in a quiet room in Scape Shoreditch, Old Street, London. By using a software named RECORDER from the Institute of Acoustics of the Chinese Academy, the speech was recorded with a microphone and stored on a computer hard disc.

4.2.3 Recording Procedures

All sentences in Mandarin characters were presented on the screen in front of the subjects. They were instructed to articulate the material at a normal speaking rate. Before the recordings, each subject was given 2 minutes to practice the name list. Speakers were instructed to repeat sentences during the recording if there was disfluency. All the stimuli were recorded in a random order, with three repetitions (in separate blocks). In total, the number of target sentences produced was 88 (sentences) × 9 (subjects) × 3 (repetitions) = 2376 tokens.

4.2.4 Annotation

Annotation at syllable level was done by using ProsodyPro (Xu, 2013), a Praat (Boersma & van Heuven, 2001) script for prosody analysis. Figure 26 shows an example of annotation. In each sentence, only the target word and the preceding syllable were labelled. This was done by marking the start of the first segment and the end of the last segment in each syllable, based on the waveform and spectrogram, with occasional reference to auditory criteria (White and Turk, 2010). Other syllables of the carrier sentences were not labelled or analysed in the current study. Prosodypro then generated all the acoustic data based on the annotation. Syllable duration was then measured at the time difference between the end point of the last segment and the start point of the first segment in each syllable.

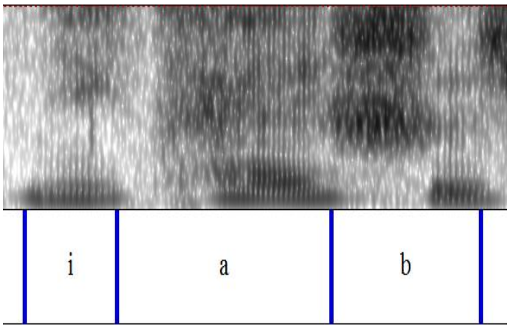


Figure 26. An example of annotation for ni+kai+shi. “i” represents the preceding syllable “ni”, “a” represents the first syllable “kai” of the target word and “b” represents the second syllable “shi.”

4.2.5 Categorization of word frequency

Frequencies of verbs were directly taken from the Modern Chinese Frequency Dictionary (Beijing Language and Culture University, 1986). Names are not available in this dictionary, so I calculated name frequency from an online name dictionary (“fotao9”, 2016) with the equation (5). The number of occurrences of each name can be found from the name dictionary. The name data is based on the identity information of 1.4 billion Chinese citizens.

$$\text{name frequency} = \text{number of occurrences} / 1.4 \text{ billion} \qquad (5)$$

Since the noun frequency and verb frequency are from different dictionaries, it is necessary to transform one of them, so it is possible to compare them. A noun that can also be found in the Modern Chinese Frequency Dictionary (Beijing Language and Culture University, 1986) is chosen as the reference. This noun is “坡, po”. The α in equation (6) is then applied to all frequencies of nouns in the study.

frequency of “坡, po” from name dictionary = α * frequency of “坡, po” from the Modern Chinese Frequency Dictionary (6)

To be able to compare the word frequency effect directly with POS effect, I transformed the gradient frequency into a categorical variable, using the median frequency as the dividing line. This way, all the target words were divided into high frequency and low frequency ones.

4.3 Results

4.3.1 Duration

As shown in Figure 27, for the POS effect, verbs are shorter than nouns in terms of mean duration. To confirm this POS effect, two one-way ANOVAs were conducted (one for monosyllabic words and the other for disyllabic words), with POS as the independent variable, duration as the dependent variable and 0.05 as the significance level. POS has a significant effect on duration, F (1, 22) = 5.069, *p* =.035 for monosyllabic words, and F (1, 62) = 12.691, *p* = 0.001 for disyllabic words. This means that nouns are significantly longer than verbs for both monosyllabic and disyllabic words.

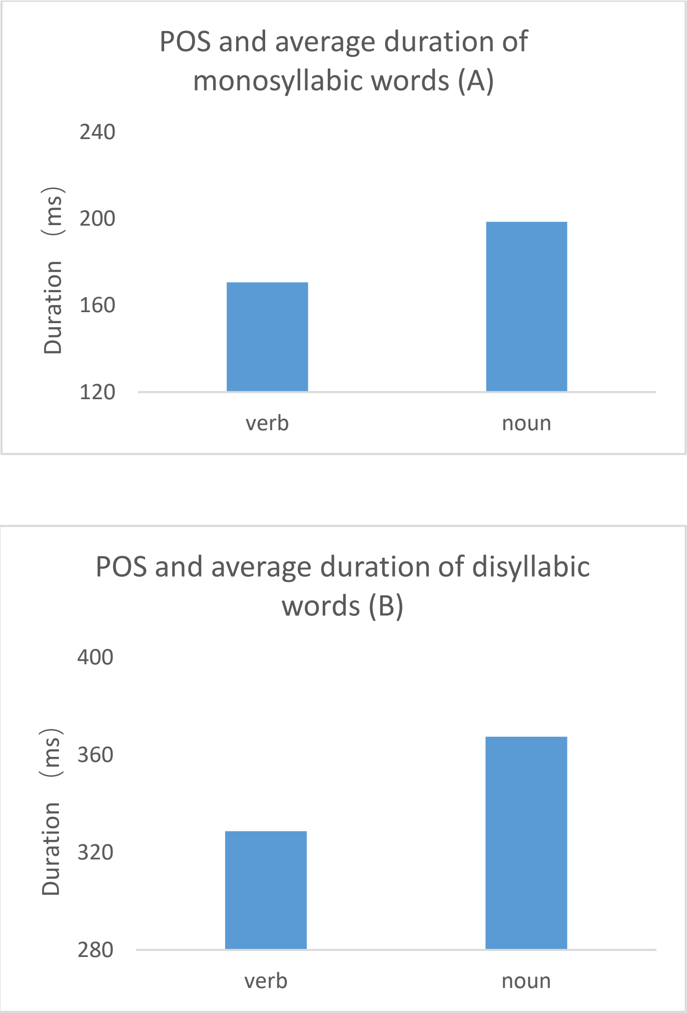


Figure 27. Part of Speech and average duration of monosyllabic words (A) and disyllabic words (B).

For the word frequency effect, the same trend on mean duration can be seen in Figure 28, with low frequency words longer than high frequency words. To confirm the frequency effect, two one-way ANOVAs were conducted (one for monosyllabic words and the other for disyllabic words), with frequency as the independent variable, duration as the dependent variable and 0.05 as the significance level. For monosyllabic words, the frequency effect is not significant, F (1, 22) =1.446, *p* =.242, while low frequency disyllabic word is significantly longer than high frequency word, F

(1, 62) = 17.477, p < 0.001. The continuous frequency is considered in a Pearson correlation analysis. The correlation between continuous frequency and duration is -.256 (p = 0.228) for monosyllabic words and -.411 (p<0.001) for disyllabic words.

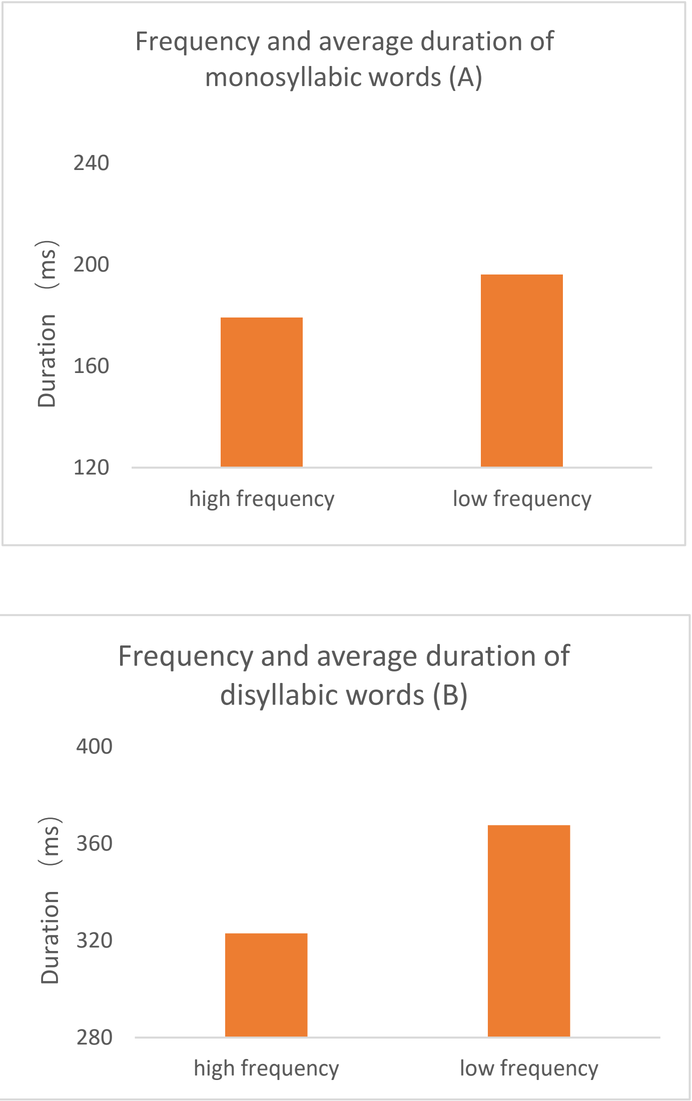


Figure 28. Frequency and average duration of monosyllabic words (A) and disyllabic words (B).

POS and frequency affect duration in the same direction, and they are related. One-way ANOVAs are not enough to compare the two independent variables, because the two variables are confounded with each other. To increase the sensitivity of the test for main effects, two ANCOVAs were conducted with POS as independent variable, frequency as covariate and duration as dependent variable for both monosyllabic words and disyllabic words. A problem is that the continuous frequencies of the target words are highly unbalanced in the data. For example, the first quantile of the continuous frequency of the target word list has 22 words and all of them are nouns. Therefore, categorised frequency instead of continuous frequency is used in the ANCOVAs analysis. The results are displayed in Table 9.

Table 9. ANCOVA results for duration of monosyllabic and disyllabic words.

Results	Monosyllabic words		Disyllabic words	
	POS	frequency	POS	frequency
F value	3.559	0.272	3.429	7.532
P value	0.073	0.608	0.069	0.008

For monosyllabic words, neither POS nor frequency has a significant effect on word duration. One possible explanation for the disappearance of the POS effect is due to the small amount of data. It is unclear if the null results are just due to a lack of power. I am cautious about the result because there are only 12 noun-verb monosyllabic word pairs here. Maybe it is indeed true that there is no POS effect on duration, but it would be better to confirm this with a larger sample size. Another factor is that the division of the words into high and low frequency ones mixed up the syllable structures, so that the frequency comparison was not based on minimal pairs of homophones. This means syllable structure, segment identity and tone can affect duration when comparing the duration of high frequency words and low frequency ones. Especially when the sample size is small (the number of target words), these effects can easily cover the frequency effect. One might suggest collecting more data. However, it is very difficult to find more verb-noun pairs in Mandarin that meet all the requirements described in Section 4.2.1 and can be embedded in meaningful sentences. Collecting more data by increasing the number of speakers cannot solve the problem that low frequency words and high frequency words are not homophones.

For disyllabic words, the POS effect is no longer significant, but the frequency effect is highly significant. This effect may have resulted from the following factors: (1) the data set is larger than monosyllabic words (32 vs. 12), (2) some high frequency words are not in homophone pairs with low frequency words because the division of the words into high and low frequency ones mixed up the syllable structures, and (3) frequency did affect duration significantly.

To rule out the second possibility, i.e., that the results were due to an uneven distribution of homophone pairs, I investigated each homophone pair of high and low frequency words. If the noun and verb from a noun-verb pair are low frequency words and are complicated in their syllable structure, which may increase duration in the same direction of low frequency, that noun-verb pair is excluded from the analysis. After the exclusion, 27 pairs were left. I tested these pairs again and the results, as shown in Table 10, is that the frequency effect is again significant with low frequency word longer than high frequency word and the POS effect is not.

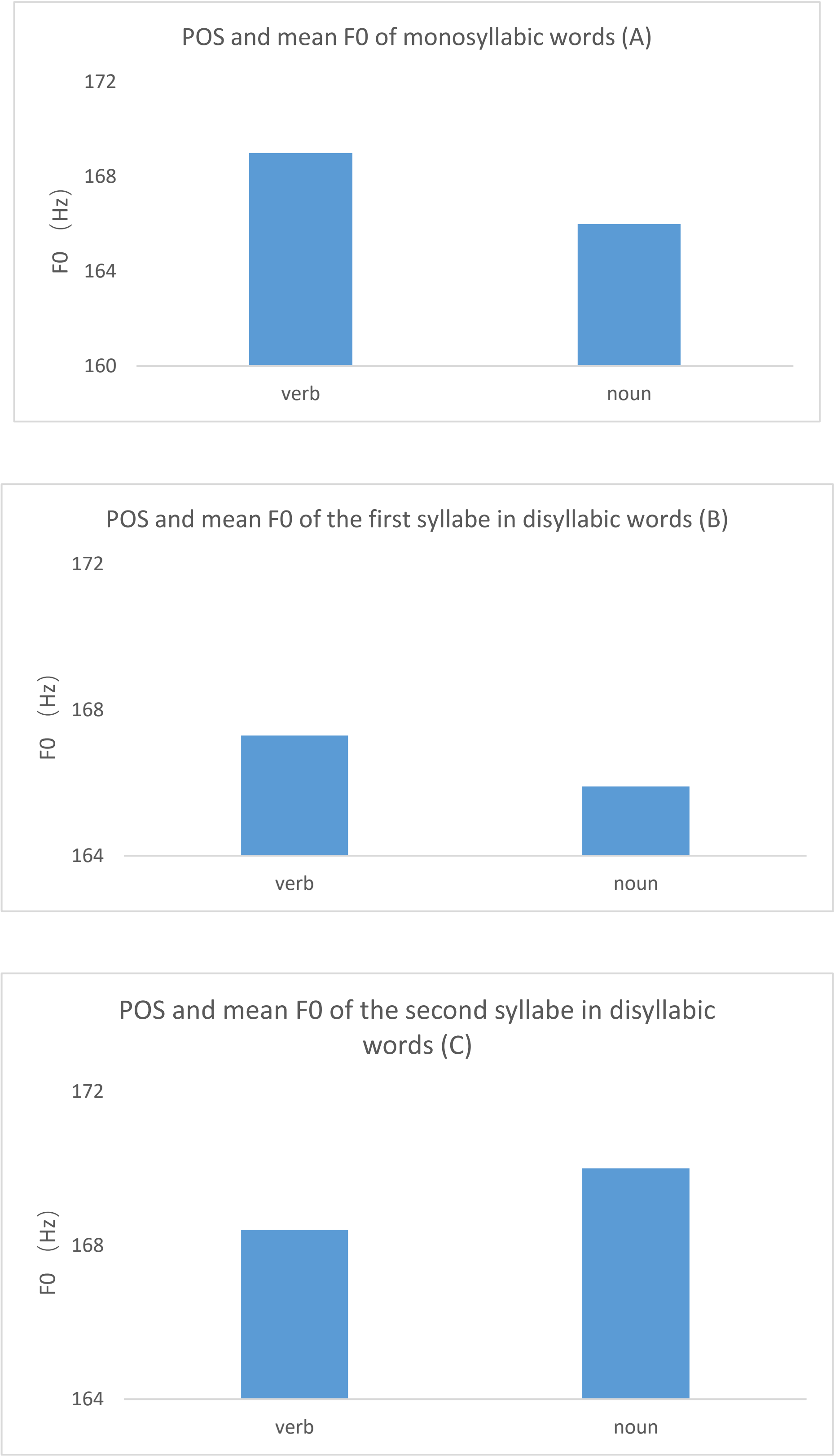
Table 10. ANCOVA results for duration of disyllabic words.

Results	POS	frequency
F value	2.286	5.241
P value	0.137	0.026

In conclusion, in this experiment, I failed to find the POS effect in either monosyllabic or disyllabic words, but the frequency effect in disyllabic words is significant with low frequency word longer than high frequency word. It is therefore possible that the frequency effect is stronger than the POS effect, so that many POS effects that have been found previously may be a result of frequency. But I am making this conclusion with caution because the dataset is small. Experiments with larger datasets will be needed in future studies.

4.3.2 F_0

Tones can affect F_0, so some pairs of sentences that do not have identical tones for target words were excluded from the data. After the exclusion, there were 12 pairs of sentences for monosyllabic words and 15 pairs for disyllabic words. A set of one-way ANOVAs was carried out with POS or frequency as the independent variable, the mean F0 of each syllable as dependent variable and 0.05 as the significance level. Figure 29 shows the mean F0 of monosyllabic words and each syllable in disyllabic words divided by POS. Figure 30 shows the mean F0 of monosyllabic words and each syllable in disyllabic words divided by categorised frequency. The results are shown in Table 11.



POS and mean F0 of monosyllabic words (A)

POS and mean F0 of the first syllabe in disyllabic words (B)

POS and mean F0 of the second syllabe in disyllabic words (C)

Figure 29. The mean F0 of monosyllabic words (A) and each syllable (B and C) in disyllabic words divided by Part of Speech.

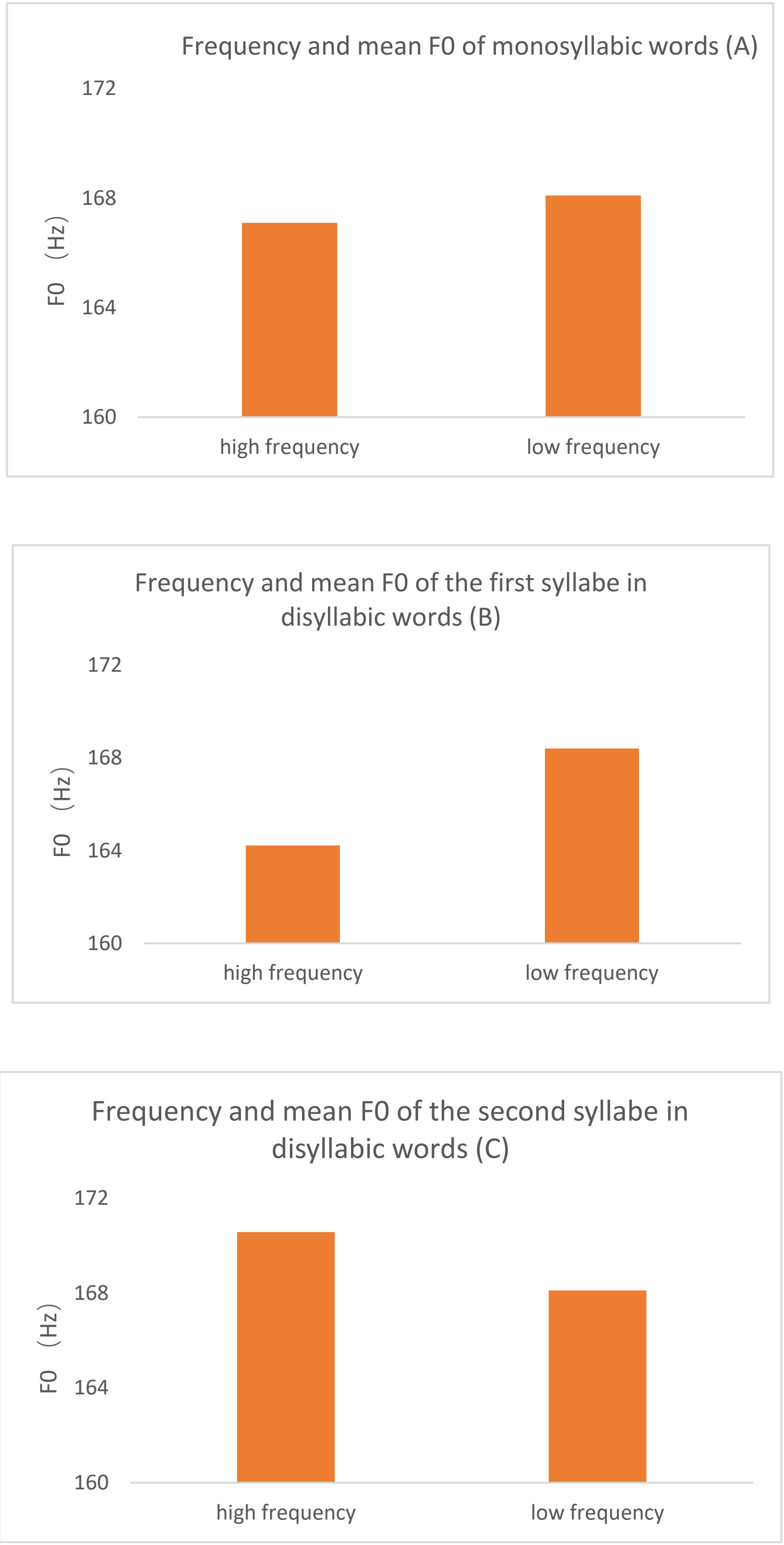


Figure 30. The mean F0 of monosyllabic words (A) and each syllable (B and C) in disyllabic words is divided by categorised frequency.

Table 11. ANOVA results for mean F0 of monosyllabic words and disyllabic words.

Levels	Monosyllabic words		Disyllabic words			
			First syllable		Second Syllable	
	POS	frequency	POS	Frequency	POS	Frequency
F value	0.171	0.021	0.077	0.733	0.002	0.316
P value	0.683	0.885	0.783	0.399	0.962	0.579

There were no POS or frequency effects on F0 in either syllable of the target words. Again, this could be because POS and frequency are related. So two ANCOVAs were performed with POS as the independent variable, frequency as the covariate and the mean F0 of each syllable as the dependent variable. As seen in Table 12, the new analysis again showed neither POS nor frequency effects on F0.

Table 12. ANCOVA results for mean F0 of monosyllabic words and disyllabic words.

Levels	Monosyllabic		Disyllabic			
			First syllable		Second Syllable	
	POS	frequency	POS	Frequency	POS	Frequency
F value	0.237	0.094	0.799	1.448	0.094	0.397
P value	0.631	0.763	0.379	0.239	0.762	0.534

4.4 Summary

This study is a preliminary test of the hypothesis that the effect of POS on duration is derived from the effect of word frequency, such that the latter is much more robust than the former. The test was performed by comparing pairs of nouns and verbs in Mandarin that are phonetically identical. Results show that there was no significant POS effect on duration or F0, but there was a significant frequency effect on duration. Given that POS does show a difference in mean duration between nouns and verbs in the present data, in the same direction as the frequency effect (tables 10 and 11), it can be concluded that this weak POS effect is derived from a more primitive effect of frequency.

5 General conclusion and Discussion

5.1 Conclusion

This dissertation investigated several timing related issues through a series of studies. I set out with three goals in mind: (1) to find out if there is a tendency towards isochrony; (2) to test whether temporal distance can be used to distinguish boundaries in different languages; and (3) to provide a clearer link between the effects of POS and word frequency on duration.

In Chapter 2, I attempted to go back to the most fundamental question about the rhythm class hypothesis, namely, is there any truth in the concept of stress timing in English, one of the few languages based on which the hypothesis was formulated in the first place? As a reference language, I also looked at Mandarin, a language that has been claimed (Lin & Wang, 2007) and confirmed (Grabe & Low, 2002; Lin & Wang, 2007; Mok & Dellwo, 2008) as syllable-timed. I replicated research from van Santen and Shih (2000) and Nakatani et al. (1981). The results of Chapter 2 showed that there is no tendency towards isochrony in English. Syllable duration was found to vary linearly both with the intrinsic duration of constituent segments (Figure 4 and Figure 5) and number of constituent segments (Figure 8). This is consistent with previous research (O’Connor, 1968; Crystal and House, 1990; van Santen and Shih, 2000). The result indicates that, in English, segments are not compressible for the sake of equal syllable duration. Inter-stress intervals, word and phrases durations were found to vary linearly with their sizes in terms of number of syllables (Figures 11-13). This is in line with previous research (Shen and Peterson, 1962; Bolinger, 1965; O'Connor, 1965; Lea, 1974; Nakatani et al., 1981). This indicates that, in English, syllables are not compressible for the sake of equal inter-stress interval or phrase duration. Without such flexibility, there is no way for syllable duration to be adjustable for showing even a tendency toward equal duration of stress groups beyond timing patterns related to linguistic functions. These results have thus dissolved the central claim of the rhythm class hypothesis, namely, that English is the epitome of a stress-timed language in which the timing of the stress groups is regulated.

In contrast, Mandarin, as an alleged syllable-timed language, showed a weak tendency toward equal word duration and phrase duration (Figures 12, 13 and 15), and this tendency was grounded on the flexibility of segment duration, which also enables a tendency toward equal syllable duration. That is, syllable duration increased at a slower rate than the increase in the intrinsic segment duration (Figures 4-6) and in the number of constituent segments (Figure 8). The results are not entirely new. The compressibility of segments can be observed in the data reported by van Santen and Shih (2000), although it was not the focus of that paper. The compressibility of syllable duration in Mandarin phrase can be seen in Xu and Wang (2009). The finding of a tendency toward equal duration of words and phrases runs counter to the classification of Mandarin as a syllable-timed language. This finding has demonstrated a further weakness of the rhythm class hypothesis. That is, even if it were weakened to the point of insisting on only a tendency toward isochrony, the search for such timing regularity may lead to the conclusion that all syllables in Mandarin are constant in duration.

In Chapter 3, the main finding is that in English, the duration of pre-boundary syllables increases constantly with break index, whereas in Mandarin, the duration increase ends after break 2. This backs up previous findings for Mandarin (Li, 1998; Yang,1997; Yang & Wang, 2002) and English (Wightman et al., 1992). In the current study, the difference between the two languages is directly shown: the duration of silent pause significantly increases beyond break level 2 in Mandarin, as if to compensate for the lack of continuous syllable lengthening, but the increase in English is less significant.

The purpose of Chapter 4 is to test the hypothesis that the effect of POS on duration is derived from the effect of word frequency, with the latter being significantly more robust than the former. The test was carried out by comparing phonetically identical pairs of Mandarin nouns and verbs. There was no significant POS influence on duration or F0, but there was a significant frequency effect on duration, according to the findings. Given that in the current data, POS does indicate a difference in mean duration between nouns and verbs in the same direction as the frequency effect (tables 10 and 11), it may be concluded that this weak POS effect is derived from a more primitive effect of frequency.

By investigating timing related issues, this dissertation provides a more comprehensive understanding of speech timing in Mandarin and English. As illustrated in Table 13, duration in English and Mandarin can be affected by intrinsic duration, stress, focus, boundary and frequency. Syllable duration varies linearly with the intrinsic duration in these two languages, but Mandarin segments are more compressible than English segments. Note that there is no equivalent of the English word stress in Mandarin, the neutral tone resembles the English weak stress (Chen and Xu, 2006; Xu and Xu, 2005), and the duration ratio of full tone to neutral tone is about 1.7:1 (Lin, 1985; Chen and Xu, 2006). In English, the stressed/unstressed duration ratio is higher: 2.18:1 (Crystal & House, 1988). As for focus, the ratio of focused to non-focused duration is 1.17:1 in Mandarin (Xu, 1999), 1.25:1 (Turk and Shattuck-Hufnagel, 2000) or 1.14:1 (Xu and Xu, 2005) in English. As for boundary, in English, pre-boundary syllable duration increases continuously with break index (Wightman et al., 1992), whereas in Mandarin, the duration increase stops beyond break 2 (Li, 1998; Yang, 1997; Yang & Wang, 2002). Frequency has an inverse relationship with duration in both English (Johnson,2004) and Mandarin. This supports previous studies on word length (Piantadosi et al., 2011; Zipf, 1935).

Table 13. Potential factors that affect duration in English and Mandarin.

Timing categories		English	Mandarin
Obligatory timing	Intrinsic duration	Not compressible	Compressible
Informational timing	stress	2.18:1	1.7:1
	focus	1.25:1 or 1.14:1	1.17:1
	boundary	Increases continuously	Ends at B2
	frequency	Inversely related	Inversely related

I have conducted a corpus study to compare the control of timing in English and Mandarin. The results of detailed duration analysis show that lexical stress in English and phrasing in both languages require clearly patterned durational cues for their marking, but important differences exist between the two languages. Once these functional factors are controlled, English segments are largely incompressible, whereas Mandarin segments are compressible. Furthermore, the duration of pre-boundary syllables in English increases linearly with break index, whereas in Mandarin, the duration increase ceases after break index 2, which is followed by the insertion of silent pauses. In addition, the last experiment shows that there is no significant POS influence on duration or F0, but there was a significant frequency effect on duration. It can be observed from these results that English and Mandarin have different ways to control timing.

5.2 Discussion

The presence and absence of compressibility found in Chapter 2 do not mean a lack of variability in segments or syllables duration in the languages examined here. Rather, they are only about whether there is additional room for segments and syllables to vary for the sake of isochrony of their enclosing units: segments for the sake of equal syllables, and syllables for the sake of equal inter-stress intervals, words or phrases. Durational variations for the sake of lexical contrasts like stress and focus are well known (Bolinger, 1972; Crystal, & House, 1988; Gussenhoven, 2008; Ladd, 1996; Nakatani et al., 1981; Turk & Shattuck-Hufnagel, 2000; van Heuven, 1994; Xu, 1999; Xu & Xu, 2005), and their contrastive nature means that they need to be sufficiently realised to guarantee intelligibility. Also known are the durational variations for the sake of marking boundaries of units like words and phrases (Lehiste, 1972; Nakatani et al., 1981; Shattuck-Hufnagel & Turk, 1996; Xu & Wang, 2009). These variations, though not necessarily categorical, serve the important function of facilitating sentence comprehension (Wagner, 2005). Functional duration variations due to lexical contrast and boundary marking are applicable to both English and Mandarin, so the differences in compressibility between the two languages found in the present study are only about the additional room for variation.

Timing resource has recently been argued to be highly valuable for speech, because speech production is likely driven by a need to maximize the rate of information transmission (Xu and Prom-On, 2019). The allocation of time resource in speech is therefore likely to be balanced between various functional needs depending on their relative importance. These needs include not only those of lexical stress and boundary marking, but also the need to guarantee intelligibility of words. The intelligibility is dependent on the identifiability of their constituent segments. And the identifiability is partially determined by the functional load of segment in the language. Functional load (Hockett, 1966; Surendran and Levow, 2004) refers to the relative importance of a phonological contrast as can be calculated based on information theory (Shannon, 1948). Other things being equal, the higher the functional load of a segment, the greater the need to guarantee its intelligibility. It is also shown that the intelligibility of a segment is related to its duration, because it takes time for articulators to move to their target positions for the segment (Birkholz et al.,2011; Lindblom, 1963; Perrier et al., 1996; Saltzman and Munhall, 1989; Xu and Prom-on, 2019), and because at normal rate, speech articulation has already reached its overall maximum speed (Xu & Prom-on, 2019). Shortening syllables and hence their constituent segments beyond certain thresholds would lead to undershoot of the articulatory targets, resulting in reduced intelligibility (Cheng & Xu, 2013). To guard against excessive shortening, it would be necessary to allocate sufficient articulation time to each segment, other things being equal. So, the lack of segmental compression in a language could arise from the need to maintain segmental intelligibility.

Interestingly, it is already shown that functional load of segments is larger in English than in Mandarin (Surendran & Levow, 2004), and this is especially true of vowels. The functional loads of consonants and vowels are 0.310 and 0.133, respectively, in English, based on Surendran and Levow's estimation, but they are 0.235 and 0.091, respectively, in Mandarin. Most interestingly, there is a likely reason for the differences in functional load between the two languages. That is, they differ vastly in the total number of different syllables. There are only 1,268 different syllables in Mandarin with tonal differences included (Yang & Xu, 1988), or about 400 possible syllables without tonal contrast or 1300 possible syllables with tones (Duanmu, 2002). In contrast, there are about 15,831 different syllables in English based on a count by Barker (2008). In other words, there are over 10 times as many syllables in English as in Mandarin syllables with tone, or nearly 40 times as many syllables in English as in Mandarin syllables without tone. To keep so many English syllables distinct from each other in speech production, it is conceivably critical that each component segment be given sufficient articulation time. In contrast, the burden of keeping only 400 Mandarin syllables distinct from each other is much lower, hence the reduced resistance to the isochrony pressure, assuming it is present. This explanation is highly speculative, of course. But it would predict that the presence and magnitude of isochrony tendency may vary across languages as a function of functional load of segments., and future research could put this to test.

Chapter 3 demonstrated clear differences between English and Mandarin in terms of the temporal marking of boundaries of various levels. In English pre-boundary syllable duration increases continuously with break index, whereas in Mandarin the duration increase stops beyond break index

2. This is consistent with previous reports for Mandarin (Yang, 1997; Li, 1998; Yang and Wang, 2002) and English (Wightman et al., 1992), respectively. Although the languages are different, they both employed different ways to mark boundaries, which are very important when encoding information. Miller (1956) found that the capacity of our short-term memory is limited, and we can only hold five to nine pieces of information in our working memories. In addition, he concluded that the threshold of short-term memory is seven chunks of information and emphasized how crucial it is to group the input sequence into chunks. That is because chunking allows us to add more information by making bigger chunks, each one with more information than before. Considering Miller's finding, boundary marking can break continuous speech into smaller chunks and allow us to encode and transmit more information.

These results in Chapter 3 cannot be attributed to speaker differences between the two corpora. Previous research with professional radio broadcaster in Mandarin also showed no significant difference in pre-boundary lengthening between minor prosodic phrase boundaries and major prosodic phrase boundaries (Li, 1998; Yang, 1997; Yang & Wang, 2002).

The frequency effect found in Chapter 4 is consistent with information theory. That is, speakers may be under a general pressure to convey as much information as possible in a given amount of time, and this pressure would lead to each word being assigned as little time as possible. But this pressure is balanced by another pressure of communication, i.e., each word also needs to be pronounced as clearly as possible, so as not to be misheard. But the chance of being misheard can be reduced by the word's predictability, which can be improved by its frequency of occurrence. So, words that are of higher frequency can afford to have less time, and thus less full articulation. Verbs, as a group, have higher frequency than nouns, and so is likely to be given less articulation time. But ultimately, it is the frequency of each individual word that partially determines its duration. This seems to have received support from the current data.

In summary, the results from all three studies are relevant to information theory. The lack of compressibility in English and the presence of it in Mandarin in Chapter 2 are probably because the functional load of segments is larger in English than in Mandarin (Surendran & Levow, 2004), and this is especially true of vowels. In Chapter 3, English and Mandarin mark major boundaries to encode and transmit more information, although they use different means. In Chapter 4, the frequency effect on duration provides further support for communication efficiency (Piantadosi et al., 2011; Shannon, 1948; Zipf, 1935). Although I have explored these topics, my research is far from exhaustive. There are some other time-related issues like speech rate, polysyllabic shortening and initial lengthening worth researching in the future, although they are not a main part of the current study.

5.3 Limitations and Future directions

In Chapter 2, I have left open the question of how the present results can address speech rhythm in general. This is because it is unclear what the definition of speech rhythm should be. It could be about what is perceived as rhythmical in speech (Arvaniti, 2012), or it could be about how continuous speech stream is broken up into groups (Cutler et al., 1997). For perceived rhythm, if it is about why a language is heard as stress timed or syllable timed, there has been a proposal that overall speech rate may be a key factor that contributes to the impression of isochrony (Dellwo & Wagner, 2003). For helping listeners break up the speech stream into groups, much research has shown that it is done mainly by marking the group boundaries with pre-boundary lengthening (Lehiste, 1972; Nakatani et al., 1981; Shattuck-Hufnagel & Turk, 1996; Wagner, 2005; Xu & Wang, 2009). This is also seen in the present data for English words (Figure 14). But surprisingly, there seems to be a lack of word final lengthening in Mandarin (Figure 15). This could mean that speech streams in Mandarin are not broken up at word boundaries, but only at the boundaries of larger units, e.g., phrases. This possibility needs to be explored in future research.

One might be concerned with the lengthening of consonants in initial position when looking at results from syllable duration at different positions within words. I examined only the syllables that occur before a B1 boundary. This means that they are not phrase final. Within words, initial, medial and final positions are considered. It has been reported that the duration of consonants is longer in word-initial position than that in word-medial position (Oller, 1973; Cooper, 1991; Fougeron and Keating, 1997; White and Turk, 2010). White, Benavides-Varela and Mády (2020) reported that lengthening of onset consonant of word-initial syllable can help English, Hungarian and Italian speakers learn an artificial language. There is evidence suggesting that initial consonant /n/ in English follows the prosodic hierarchy, with greater initial lengthening after larger boundaries (Fougeron and Keating, 1997). It would be ideal to consider initial consonant lengthening, but I cannot guarantee the measurement of initial consonant duration is correct, especially the measurement of stops. This is because the annotation of the close of stops in some samples are not reliable, as it is difficult to annotate the starting point of closure. Due to the infeasibility, initial consonant lengthening is not addressed in the current analysis. Although initial lengthening is not considered in my analysis, my results (Figure 14-15) are in line with previous research (Nakatani et al., 1981; Xu and Wang, 2009). Since it is a study comparing two languages, if initial lengthening has an effect on the analysis, it is reasonable to assume it would have affected the analysis in both languages. It is unlikely that the difference between two languages can be overturned.

Another interesting issue to take into consideration is word length. White and Turk (2010) investigated polysyllabic shortening in British English by measuring the duration of stressed syllable in monosyllabic, disyllabic and trisyllabic words. They discovered that there is strong evidence of polysyllabic shortening in both left-headed and right-headed words in accented words. While in unaccented words, polysyllabic shortening is rare or absent. Therefore, both accent and word length are important factors that affect duration. It would be ideal to consider accent and word length in my analysis. They put their keywords in sentences designed to be read as meaningful utterances. By putting certain words in block capitals and telling participants to emphasise those capitalised words, they managed to control the location of pitch accent. They also excluded utterances if the keyword was not accented in the unaccented condition or unaccented in the accented condition. However, in my study, I used the Boston University Radio News Corpus. The ToBI system was used in the prosodic labelling in this corpus. As reported by Wightman (2002), the annotation of pitch accent is not very reliable in ToBI. Labellers are required to label any syllable that sounds perceptually prominent to be “pitch-accented”, regardless of the absence of tonal marking of pitch accent (Beckman and Ayers, 1997). Thus, the reliability of pitch accent annotation in the Boston University Radio News Corpus is questionable, as the ToBI system was used in the prosodic labelling in this corpus. For this reason, I did not include word length and pitch accent in my analysis. However, these factors can be considered in future research. The inclusion of mixed-effects regression models should also be good to consider in future study.

It is critical to acknowledge that the segmentation times and phone durations are provided in units of 10 milliseconds in the English corpus. This can affect the durational analysis and results and should be investigated in future research. Another issue is that the phonetic alignments for the news recorded in the laboratory in the Boston University Radio News Corpus were hand-corrected, while those recorded during broadcast were not. Whether this affects the results is something worth looking at in future research.

In Chapter 3, a potential confounding factor when comparing English and Mandarin is the different criteria used in the labelling of the break indices between ToBI and C-ToBI. As mentioned in the background, the determination of break index in English depends heavily on intonation annotation (Beckman & Ayers, 1997.). Critically, break index 3 is obligatory whenever a phrase accent is present, which by definition marks the end of an intermediate phrase even if there is no silent pause. The virtual overlap of temporal distance with break index 3 in Figure 22 shows that, indeed, little silence accompanied this break level. However, despite the lack of silence at break index 3 in the English corpus, significant pre-boundary lengthening was found. This indicates that English syllables are much more flexible than Mandarin in terms of lengthening beyond break index 2. On the other hand, despite the robust difference, cross-boundary temporal distance, consisting of durations of both pre-boundary syllable and silent pause, seems to be a common marker of boundary strength in both languages.

One would ask why there is a difference between English and Mandarin in terms of pre-boundary lengthening. For now, I do not have an answer. Maybe it is necessary to look at lengthening in other languages. For example, the durational making of both domain edges and domain heads is reduced in Spanish (Frota et al., 2007; Ortega-Llebaria and Prieto, 2007), as is noted in White and Mattys (2007). This phenomenon could be historical. Whether this is true is an empirical question that can be answered only after an investigation of many more languages.

Although the analysis did not include initial lengthening, it is reported that the initial consonant /n/ in English follows the prosodic hierarchy, with greater initial lengthening after larger boundaries (Fougeron and Keating, 1997). Initial consonants in Mandarin show the same trend (Cao, 2005). It will be interesting to include initial lengthening in future research.

Word length is another important factor that is not included in the Chapter 3 analysis and worth looking at in the future. White (2002) found that stressed syllable codas in penultimate and antepenultimate position are affected by final lengthening. It would be ideal to consider word length, stress location and syllable structure in the analysis. However, the study would suffer from data sparsity and intrinsic duration of segments and number of segments are likely to override other effects if the data were divided more based on word length, stress location and syllable structure after considering break index, stress and speaker. This is because the corpus was not designed for this study. It will be interesting to include these factors in the analysis of a larger corpus. Word length is also related to polysyllabic shortening. As mentioned earlier, it is found that there is strong evidence of polysyllabic shortening in both left-headed and right-headed words in accented words. But in unaccented words, polysyllabic shortening is minimal or absent (White and Turk, 2010). But data sparsity will still be a problem if I include word length and accent in the current analysis. Hopefully, these factors can be considered in future study.

Chapter 4 is a study based on a limited size of data. I am making my conclusion cautiously. It is necessary to test if frequency effect is more robust than POS effect in a bigger data set as well as in other languages. In future research, it is also possible to conduct a parallel study in English. In addition, it might be better to analyse the data using a mixed-effects regression analysis, with random effects of speaker and item.

5.4 Concluding remarks

In this thesis, a corpus study was conducted to compare how timing is controlled in English and Mandarin. The results shows that these two languages have different ways to control timing. When functional factors like stress, focus, boundary are controlled for, English segments and syllables are largely incompressible, while Mandarin segments and syllables demonstrate compressibility. Therefore, this thesis presents evidence that there is no tendency toward isochrony in English, but that Mandarin shows a weak tendency toward isochrony. In addition, the duration of pre-boundary syllables in English increases linearly with break index. However, in Mandarin, the increase in duration ceases after break index 2, which is followed by the insertion of silent pause. Since this demonstrates a significant difference of these two languages in terms of boundary marking, the results can be directly applied to second language learning. Furthermore, the last study demonstrates that there is no statistically significant influence of POS on duration or F0 in Mandarin. However, a significant frequency effect was observed on duration and effects of POS are likely a by-product of word frequency. This can facilitate our understanding of the rationale behind considering POS as an essential input feature in the training process in both speech recognition and speech synthesis. Needless to say, the work in this thesis is far from complete. Many factors like initial lengthening of consonants and word length can be considered in a bigger data set in future study.

6 Bibliography

Abercrombie, D. (1964). A phonetician's view of verse structure. *Linguistics, 2*(6), 5-13. doi:10.1515/ling.1964.2.6.5

Abercrombie, D. (1964). "Syllable quantity and enclitics in English" *in Honour of Daniel Jones*. eds. D. Abercrombie, D. Fry, D. MacCarthy, N. Scott, and J. Trim (London: Longman), 216–222.

Abercrombie, D. (1967). Elements of general phonetics. Edinburgh: Edinburgh University Press.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica* 66, 46–63.

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of phonetics, 40*(3), 351-373. doi:10.1016/j.wocn.2012.02.003

Arvaniti, A., & Rodriquez, T. (2013). The role of rhythm class, speaking rate, and F 0 in language discrimination. *Laboratory Phonology*, 4(1), 7-38.

Aylett, M., & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration on spontaneous speech. *Language And Speech, 47*, 31-56.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *J. Acoust. Soc. Am., 119*(5), 3048-3058. doi:10.1121/1.2188331

Barker, C. (2008). How Many Syllables Does English Have? Available at: https://web.archive.org/web/20160923005626/http://semarch.linguistics.fas.nyu.edu:80/barker/Syllables/index.txt (Accessed April 25, 2021).

Beckman, M. E., & Ayers, G. (1997). Guidelines for ToBI labelling. *The OSU* Research Foundation, 3, 30.

Beckman, M. E. (1992). Evidence for speech rhythms across languages. *Speech perception, production and linguistic structure*, 457-463.

Beijing Language and Culture University. (1986), “Modern Chinese Frequency Dictionary”.

Bertinetto, P. M. (1989). Reflections on the dichotomy ‘stress’ vs.‘syllable-timing’. *Revue de phonétique appliquée, 91*(93), 99-130.

Bertinetto, P. M., & Bertini, C. (2008). On modeling the rhythm of natural languages*. Paper presented at the Proceedings of the Fourth International Conference on Speech Prosody*.

Birkholz, P., Kroger, B. J., & Neuschaefer-Rube, C. (2011). Model-based reproduction of articulatory trajectories for consonant-vowel sequences. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(5), 1422-1433. doi:10.1109/TASL.2010.2091632

Birkholz, P., Kroger, B. J., & Neuschaefer-Rube, C. (2010). Model-based reproduction of articulatory trajectories for consonant–vowel sequences. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5), 1422-1433.

Bloch, B. (1950). Studies in Colloquial Japanese IV Phonemics. *Language, 26*(1), 86-125. doi:10.2307/410409

Boersma, P. & Weenink, D. (2011). Praat: Doing phonetics by computer. (Version 5.2.36) [Computer Program]. Available at: ⟨http://www.fon.hum.uva.nl/praat/⟩.

Bolinger, D. L. (1965). Forms of English: Accent, morpheme, order: Harvard University Press.

Bolinger, D. L. (1972). Accent is predictable (if you’re a mind reader). Language, 48, 633–644.

Borzone de Manrique, A. M., & Signorini, A. (1983). Segmental duration and rhythm in Spanish. *Journal of Phonetics, 11*(2), 117-128.

Bulut, M., Lee, S., & Narayanan, S. S. (2007). Analysis of emotional speech prosody in terms of part of speech tags. In INTERSPEECH (pp. 626-629).

Campbell, W. N., & Isard, S. D. (1991). Segment durations in a syllable frame. *Journal of phonetics*, 19(1), 37-47.

Cao, J. (2005). Types of segmental lengthening and their prosodic functions. *Journal of School of Chinese Language and Culture Nanjing Normal University*, 4, 160-167.

Chen, Y., & Xu, Y. (2006). Production of weak elements in speech–evidence from F_0 patterns of neutral tone in Standard Chinese. Phonetica, 63(1), 47-75.

Cheng, C., & Xu, Y. (2013). Articulatory limit and extreme segmental reduction in Taiwan Mandarin. *The Journal of the Acoustical Society of America, 134*(6), 4481. doi:10.1121/1.4824930

Choi, J. Y., Hasegawa-Johnson, M., & Cole, J. (2005). Finding intonational boundaries using acoustic cues related to the voice source. *The Journal of the Acoustical Society of America*, 118(4), 2579-2587.

Chomsky, N. (2002). An interview on minimalism. N. Chomsky, *On Nature and Language*, 92–161.

Classe, A. (1939). The rhythm of English prose. B. Blackwell.

Coker, C., Umeda, N., & Browman, C. (1973). Automatic synthesis from ordinary English test. *IEEE Transactions on Audio and Electroacoustics, 21*(3), 293-298. doi:10.1109/TAU.1973.1162458

Cooper, A. M. (1991). Laryngeal and oral gestures in English/p, t, k. *In Proceedings of the 12th international congress of phonetic sciences* (Vol. 2, pp. 50-53).

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Syllabic stress. *The Journal of the Acoustical Society of America*, 83(4), 1574-1585.

Crystal, T. H., & House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *The Journal of the Acoustical Society of America*, 88(1), 101-112.

Cutler, A., Dahan, D., and Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and speech* 40, 141–201.

Dankovičová, J., & Dellwo, V. (2007). Czech speech rhythm and the rhythm class hypothesis. *Paper presented at the Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of phonetics, 11*, 51-62.

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for ∆C. *Language and language-processing*, 231-241.

Dellwo, V., and Wagner, P. (2003). “Relations between language rhythm and speech rate,” *in Proceedings of the 15th International Congress* of Phonetic Sciences, Barcelona. 471–474.

De Jong, K. J. (2001). Rate-induced resyllabification revisited. *Language and Speech*, 44(2), 197-216.

Deterding, D. (2001). The measurement of rhythm: a comparison of Singapore and British English. *Journal of phonetics, 29*(2), 217-230. doi:10.1006/jpho.2001.0138

Dellwo, V., Leemann, A., & Kolly, M. J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, 137(3), 1513-1528.

Duanmu, S. (2000). The phonology of Standard Chinese. Oxford: Oxford University Press.

Eriksson, A. (1991). Aspects of Swedish speech rhythm. Sweden: Department of Linguistics, University of Göteborg.

Fletcher, J., Hardcastle, W. J., Laver, J., & Gibbon, F. E. (2010). The handbook of phonetic sciences. *In The prosody of speech: timing and rhythm* (pp. 521-602). Oxford: Blackwell.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The journal of the acoustical society of* America, 101(6), 3728-3740.

Frota, S., D Imperio, M., Elordieta, G., Prieto, P., & Vigário, M. (2007). The phonetics and phonology of intonational phrasing in Romance. *Amsterdam Studies in the Theory and History of Linguistic Science Series* 4, 282, 131.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America, 27*(4), 765-768. doi:10.1121/1.1908022

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, 1(2), 126-152.

Gao, H., and Xu, Y. (2010). “Ambisyllabicity in English: How Real is it?,” *in Proceedings of the 9th Phonetics Conference of China (PCC2010),* Tianjin, China.

Gibbon, D. (2003). Computational modelling of rhythm as alternation, iteration and hierarchy*. Paper presented at the Proceedings of ICPhS*.

Grabe, E., and Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers Lab*. Phonol. 7, 515–546. doi: 10.1515/9783110197105.2.515

Gussenhoven, C. (2008). Types of focus in English. In Topic and focus (pp. 83-100). Springer, Dordrecht.

Hockett, C. F. (1966). The quantification of functional load-a linguistic problem.

Hoequist, J. C. (1983). Durational Correlates of Linguistic Rhythm Categories. *Phonetica, 40*(1), 19-31. doi:10.1159/000261679

Hollien, H., and Majewski, W. (1977). Speaker identification by long-term spectra under normal and distorted speech conditions. J. Acoust. Soc. Am. 62, 975–980. doi: 10.1121/1.381592

Johnson, K. (2004). *Massive reduction in conversational American English.* Paper presented at the Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium.

Kelso, J. A. S., Saltzman, E. L., and Tuller, B. (1986). The dynamical perspective on speech production: data and theory. *J. Phon*. 14, 29–59. doi: 10.1016/S0095-4470(19)30608-4

Kimball, O., Ostendorf, M., & Bechwati, I. (1992). Context modeling with the stochastic segment model. *IEEE Transactions on Signal Processing, 40*(6), 1584-1587. doi:10.1109/78.139267

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of phonetics*, 3(3), 129-140.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *J. Acoust. Soc. Am*. 59, 1208–1221. doi: 10.1121/1.380986

Knight, R. (2011). Assessing the temporal reliability of rhythm metrics. *Journal of the International Phonetic Association, 41*(3), 271-281. doi:10.1017/S0025100311000326

Kohler, K. J. (2009). Rhythm in Speech and Language. *Phonetica, 66*(1-2), 29-45. doi:10.1159/000208929

Ladd, D. R. (1988). Declination ‘‘reset’’and the hierarchical organization of utterances. The Journal of the Acoustical Society of America, 84(2), 530-544.

Ladd, D. R. (1996). Intonational phonology. Cambridge: Cambridge University Press.

Ladefoged, P. (1975). A course in phonetics. New York: Harcourt Brace Jovanovich.

Lea, W. A. (1974). Prosodic Aids to Speech Recognition*. 4. A General Strategy for Prosodically-Guided Speech Understanding*.

Lea, W. A. (1975). Isochrony and disjuncture as aids to syntactic and phonological analysis. *The Journal of the Acoustical Society of America, 57*(S1), S33-S33. doi:10.1121/1.1995181

Lea, W. A. (1980). Trends in speech recognition: Prentice Hall PTR.

Lehiste, I. (1972). The Timing of Utterances and Linguistic Boundaries. *The Journal of the Acoustical Society of America, 51*(6B), 2018-2024. doi:10.1121/1.1913062

Lehiste, I. (1977). Isochrony reconsidered. *Journal of phonetics*.

Lehiste, I. (1979). Perception of sentence and paragraph boundaries. *Frontiers of speech communication research*, 191-201.

Li, A. (1998). An analysis of Mandarin prosodic phrase duration. http://ling.cass.cn/yuyin/report/report_1998.htm

Li, A. (2002). Chinese prosody and prosodic labeling of spontaneous speech*. Paper presented at the Speech Prosody 2002*, International Conference.

Li, A., Chen, X., Sun, G., Hua, W., Yin, Z., Zu, Y., Zheng, F., & Song, Z. (2000). The phonetic labeling on read and spontaneous discourse corpora*.* Paper presented at the Sixth International Conference on Spoken Language Processing.

Li, A., Lin, M., Chen, X., Zu, Y., Sun, G., Hua, W., et al. (2000b). “Speech corpus of Chinese discourse and the phonetic research,” in Paper Presented at the ICSLP’2000.

Lin, H., & Wang, Q. (2007). Mandarin rhythm: An acoustic study. *Journal of Chinese language computing, 17*(3), 127-140.

Lin, T. (1985). Preliminary experiments on the nature of Mandarin neutral tone. Working Papers in Experimental Phonetics. Beijing University Press, Beijing, 1-26.

Lindblom, B. (1963). Spectrographic Study of Vowel Reduction. *The Journal of the Acoustical Society of America, 35*(11), 1773-1781. doi:10.1121/1.1918816

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. *In Speech production and speech modelling* (pp. 403-439). Dordrecht: Springer Netherlands.

Loukina, A., Kochanski, G., Rosner, B., Keane, E., & Shih, C. (2011). Rhythm measures and dimensions of durational variation in speech. *The Journal of the Acoustical Society of America*, 129(5), 3258-3270.

Lyons, J. W. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus. *National Institute of Standards and Technology.*

Martin, J. G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological review, 79*(6), 487. doi:10.1037/h0033467

Meunier, C., & Espesser, R. (2011). Vowel reduction in conversational speech in French: The role of lexical factors. *Journal of phonetics, 39*(3), 271-278. doi:10.1016/j.wocn.2010.11.008

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2), 81.

Miller, M. (1984). On the perception of rhythm. *Journal of Phonetics*, 12(1), 75-83.

Mok, P. (2009). On the syllable-timing of Cantonese and Beijing Mandarin. *Chinese Journal of Phonetics, 2*, 148-154.

Mok, P., & Dellwo, V. (2008). *Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing*

Mandarin and English. Paper presented at the Proceedings of Speech Prosody.

Names. (2016). In fotao9.com. Retrieved from http://www.sosuo.name/tong/.

Nakatani, L. H., O'Connor, K. D., & Aston, C. H. (1981). Prosodic aspects of American English speech rhythm. *The Journal of the Acoustical Society of America, 69*(S1), S82-S82. doi:10.1121/1.386084

Nolan, F., & Asu, E. L. (2009). The Pairwise Variability Index and Coexisting Rhythms in Language. *Phonetica, 66*(1-2), 64-77. doi:10.1159/000208931

Nolan, F., & Jeon, H.-S. (2014). Speech rhythm: a metaphor? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 369*(1658), 20130396. doi:10.1098/rstb.2013.0396

O'Connor, J. D. (1965). The perception of time intervals. Progress Report 2, *London: Phonetics Laboratory*, University College.

O’Connor, J. D. (1968). The duration of the foot in relation to the number of component sound-segments. *Progress Report*, 3, 1-6.

Oller, D. K. (1973). The effect of position in utterance on speech segment duration in English. *The journal of the Acoustical Society of America*, 54(5), 1235-1247.

O'Malley, M., Kloker, D., and Dara-Abrams, B. (1973). Recovering parentheses from spoken algebraic expressions. IEEE Trans. Audio Electroacoust. 21, 217–220. doi: 10.1109/TAU.1973.1162449

O’Rourke, E. (2008). Speech rhythm variation in dialects of Spanish: applying the pairwise variability index and variation coefficients to Peruvian Spanish*. Paper presented at the Proc. Fourth Conf. on* Speech Prosody.

Ortega-Llebaria, M., & Prieto, P. (2007). Disentangling stress from accent in Spanish: Production patterns of the stress contrast in deaccented syllables. *Amsterdam Studies in the Theory and History of Linguistic Science Series* 4, 282, 155.

Ostendorf, M., Price, P. J., & Shattuck-Hufnagel, S. (1995). The Boston University radio news corpus. Linguistic Data Consortium, 1-19.

Perrier, P., Ostry, D. J., & Laboissière, R. (1996). The equilibrium point hypothesis and its application to speech motor control. *Journal of* speech and hearing research, 39(2), 365-378.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9), 3526-3529.

Pike, K. L. (1945). The Intonation of American English. Michigan: University Press.

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *J. Acoust. Soc. Am., 118*(4), 2561-2569. doi:10.1121/1.2011150

Pointon, G. E. (1980). Is Spanish Really Syllable-Timed? *Journal of phonetics, 8*(3), 293-304.

Port, R. F., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in Japanese. *The Journal of the Acoustical Society of America, 81*(5), 1574. doi:10.1121/1.394510

Qian, Y., Chu M., & Pan, W. (2001). The acoustic analysis of prosodic boundaries in Mandarin. *The Proceeding of 5th National Conference on Modern Phonetics*.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73*(3), 265-292. doi:10.1016/S0010-0277(99)00058-X

Roach, P. (1982). On the distinction between ‘stress-timed’ and ‘syllable-timed’ languages. *Linguistic controversies, 73*, 79.

Rubach, J., & Booij, G. E. (1985). A grid theory of stress in polish. *Lingua, 66*(4), 281-320. doi:10.1016/0024-3841(85)90032-4

Saltzman, E. L., & Munhall, K. G. (1989). A Dynamical Approach to Gestural Patterning in Speech Production. *Ecological Psychology, 1*(4), 333-382. doi:10.1207/s15326969eco0104_2

Schafer, A. J., Speer, S. R., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of psycholinguistic research*, 29, 169-182.

Scott, D. R. (1982). Duration as a cue to the perception of a phrase boundary a. *Journal of the Acoustical Society of America, 71*(4), 996-1007. doi:10.1121/1.387581

Scott, D. R., Isard, S. D., & de Boysson-Bardies, B. (1985). Perceptual isochrony in English and in French. *Journal of Phonetics*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379-423. doi:10.1002/j.1538-7305.1948.tb01338.x

Shattuck-Hufnagel, S., & Turk, A. (1996). A prosody tutorial for investigators of auditory sentence processing. *J Psycholinguist Res, 25*(2), 193-247. doi:10.1007/BF01708572

Shen, X. N. S. (1990). the prosody of Mandarin Chinese (Vol. 118). Univ of California Press.

Shen, Y., & Peterson, G. G. (1962). *Isochronism in English*. Buffalo: Department of Anthropology and Linguistics, University of Buffalo.

Shih, C. (2000). A declination model of Mandarin Chinese. In Intonation (pp. 243-268). Springer, Dordrecht.

Steever, S. B. (1987). Tamil and the Dravidian languages. *The world's major languages*, 725-746.

Steven, T. P., Harry, T., & Edward, G. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences, 108*(9), 3526. doi:10.1073/pnas.1012551108

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., . . . Hirschberg, J. (1992). *ToBI: A standard for labeling English prosody.* Paper presented at the Second international conference on spoken language processing.

Sorensen, J. M., Cooper, W. E., & Paccia, J. M. (1978). Speech timing of grammatical categories. *Cognition, 6*(2), 135-153. doi:10.1016/0010-0277(78)90019-7

Surendran, D., & Levow, G. A. (2004). The functional load of tone in Mandarin is as high as that of vowels. *In Speech Prosody 2004*, International Conference.

Sun, X. (2002). Pitch accent prediction using ensemble machine learning. *In Seventh international conference on spoken language processing.*

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *The* Journal *of the Acoustical Society of America, 101*(1), 514. doi:10.1121/1.418114

Tiffany, W. R. (1980). The effects of syllable structure on diadochokinetic and reading rates. *Journal of Speech, Language, and Hearing Research*, 23(4), 894-908.

Turk, A.E., Shattuck-Hufnagel, S., (2000). Word-boundary-related duration patterns in English. Journal of Phonetics 28, 397–440.

Uldall, E. T. (1971). Isochronous stresses in RP. *Form and substance*, 205-201.

Uldall, E. T. (1978). Rhythm in very rapid RP. *Lang. Speech* 21, 397–402. doi: 10.1177/002383097802100415

van Heuven, V. J. (1994). What is the smallest prosodic domain? In P. A. Keating (Ed.), Papers in laboratory phonology (pp. 76–98). Cambridge: Cambridge University Press.

van Santen, J. P., & Shih, C. (2000). Suprasegmental and segmental timing models in Mandarin Chinese and American English. *The Journal of the Acoustical Society of America, 107*(2), 1012. doi:10.1121/1.428281

Wagner, M. (2005). *Prosody and recursion.* Massachusetts Institute of Technology, Massachusetts. Retrieved from https://dspace.mit.edu/handle/1721.1/33713

Wang, B., Xu, Y., & Ding, Q. (2017). Interactive Prosodic Marking of Focus, Boundary and Newness in Mandarin. Phonetica, 75(1), 24-56.

Warner, N., & Arai, T. (2001). The role of the mora in the timing of spontaneous Japanese speech. *The Journal of the Acoustical Society of America*, 109(3), 1144-1156.

Wenk, B. J., & Wioland, F. (1982). Is French really syllable-timed? *Journal of phonetics*.

White, L. S. (2002). English speech timing: a domain and locus approach (Doctoral dissertation, University of Edinburgh).

White, L. (2014). Communicative function and prosodic form in speech timing. *Speech communication*, 63, 38-54.

White, L., Benavides-Varela, S., & Mády, K. (2020). Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues?. *Journal of Phonetics*, 81, 100982.

White, L., Delle Luche, C., & Floccia, C. (2016). Five-month-old infants' discrimination of unfamiliar languages does not accord with "rhythm class". *In Proceedings of Speech Prosody* (pp. 567-571).

White, L., and Malisz, Z. (2020). "Speech rhythm and timing," *in The Oxford Handbook of Language Prosody*. eds. A. Gussenhoven and A. Chen (Oxford: Oxford University Press), 167–182.

White, L., & Turk, A. E. (2010). English words on the Procrustean bed: Polysyllabic shortening reconsidered. *Journal of Phonetics*, 38(3), 459-471.

White, L., Mattys, S. L., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language*, 66(4), 665-679.

Wightman, C. W. (2002). ToBI or not ToBI?. *In Speech Prosody* 2002, International Conference.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America, 91*(3), 1707-1717. doi:10.1121/1.402450

Xiong, Z. (2003). An Acoustic Study of the Boundary Features of Prosodic Units. *Applied Linguistics, 2*, 116-121.

Xu, Y., 1999. Effects of tone and focus on the formation and alignment of F0 contours. Journal of Phonetics 27, 55–105.

Xu, Y. (2009). Timing and coordination in tone and intonation-An articulatory-functional perspective. *Lingua, 119*(6), 906-927. doi:10.1016/j.lingua.2007.09.015

Xu, Y. (2013). ProsodyPro: A tool for large-scale systematic prosody analysis. *In Proceedings of Tools and Resources for the Analysis of Speech Prosody* (TRASP 2013) (pp. 7–10). Aixen-Provence, France.

Xu, Y. (2019). Prosody, tone, and intonation. In The Routledge handbook of phonetics (pp. 314-356). Routledge.

Xu, Y., & Prom-On, S. (2019). Economy of effort or maximum rate of information? Exploring basic principles of articulatory dynamics. Frontiers in psychology, 10, 2469.

Xu, Y., & Wang, M. (2009). Organizing syllables into groups—Evidence from F0 and duration patterns in Mandarin. *Journal of phonetics, 37*(4), 502-520. doi:10.1016/j.wocn.2009.08.003

Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. Journal of Phonetics, 33(2), 159-197.

Yang, Y. (1997). Prosodic cues to syntactic boundaries. *Acta Acustica*, 22 (5), 414-421.

Yang, Y., & Wang, B. (2002). *Acoustic correlates of hierarchical prosodic boundary in Mandarin.* Paper presented at the Speech Prosody 2002, International Conference.

Yang, S.-A., and Xu, Y. (1988). An acoustic-phonetic oriented system for synthesizing Chinese. *Speech Comm* 7, 317–325.

Yin, H. (2007). Serial verb constructions in English and Chinese. *In Proceedings of the 2007 Annual Conference of the Canadian Linguistic Association* (Vol. 17).

Yuan, J., & Liberman, M. (2015). Investigating consonant reduction in Mandarin Chinese with improved forced alignment. *In Sixteenth annual conference of the international speech communication association*.

Zhao, Y., & Jurafsky, D. (2009). The effect of lexical frequency and Lombard reflex on tone hyperarticulation. Journal of Phonetics, 37(2), 231-247.

Zipf, G. K. (1935). The psychobiology of language. Cambridge MA: MIT Press.

7 Appendix 1: sentence list

Words	Sentences
1.要/耀	快抓绳子，牛要（v.）逃跑了。
yao4	有人劫狱，牛耀（n.）逃跑了。
2.能/能	平时有空的时候要多阅读，书能（v.）带给我们很多信息。
neng2	为了让我们在股市赚到钱，舒能（n.）带给我们很多信息。
3.会/惠	没饲料的时候，猪会（v.）吃草。
hui4	减肥那段时间，朱惠（n.）吃草。
4.想/响	我知道，你想（v.）去上海了。
xiang3	别找了，倪响（n.）去上海了。
5.肯/肯	这件事不难办，你肯（v.）出面就能解决。
ken3	这件事不难办，尼肯（n.）出面就能解决。
6.许/栩	这里是公共场合，不许（v.）抽烟。
xu3	为缓解生活压力，杜栩（n.）抽烟。
7.准/准	这里是公共场合，不准（v.）抽烟。
zhun3	为缓解失恋痛苦，杜准（n.）抽烟。
8.爱/艾	这条街的邻居都知道，你爱（v.）画画。
ai4	为了表达自己的想法，李艾（n.）画画。
9.嫌/贤	最近隔壁在装修，你嫌（v.）白天太吵了。
xian2	大家跟我投诉说：李贤（n.）白天太吵了。
10.怕/坡	种种迹象表明：你怕（v.）寂寞。

pa4/po1	种种迹象表明：李坡（n.）寂寞。
11.看/叹	病人血压正常，你看（v.）没什么大问题。
kan4/tan4	生活背景单纯，李叹（n.）没什么大问题。
12.说/硕	咱们早就约好了，你说（v.）明天去吃火锅。
shuo1/shuo4	刚刚电话确认了，李硕（n.）明天去吃火锅。
13.不用/杜友	也就几百米，不用（v.）坐车。
bu2yong4/du4you3	为了赶时间，杜友（n.）坐车。
14.敢于/甘雨	你这人性格不错，敢于（v.）调侃自己。
gan3yu2/gan1yu3	为活跃现场气氛，甘雨（n.）调侃自己。
15.乐意/陆毅	真正大方的人，乐意（v.）和别人分享自己的午餐。
le4yi4/lu4yi1	为了不剩下饭，陆毅（n.）和别人分享自己的午餐。
16.需要/许耀	要想成功，需要（v.）不断尝试。
xu1yao4/xu3yao4	为了成功，许耀（n.）不断尝试。
17.必须/毕虚	要想成为天下第一，必须（v.）苦练内功。
bi4xu1	为了成为天下第一，毕虚（n.）苦练内功。
18.善于/山雨	这条街的邻居都知道，你善于（v.）踢足球。
shan4yu2/shan1yu3	为了强化腿部的骨骼，倪山雨（n.）踢足球。
19.应当/应刚	如果要在前方变更车道，应当（v.）仔细观察。
ying1dang1/ying1gang1	为了从血迹中找到线索，应刚（n.）仔细观察。

20.可能/柯能	你呆着不舒服，可能（v.）想走。
ke3neng2/ke1neng2	这里环境不好，柯能（n.）想走。
21.打算/杜双	刘云买了双运动鞋，打算（v.）去健身房瘦身。
da3suan4/du4shuang1	为了买一条新裙子，杜双（n.）去健身房瘦身。
22.开始/开使	为了提高身体素质，你开始（v.）学习游泳。
Kai1shi3	为了提高身体素质，李开使（n.）学习游泳。
23.装作/庄作	刘云怕惹祸上身，装作（v.）没看见。
zhuang1zuo4	地上路标不明显，庄作（n.）没看见。
24.认为/任伟	张涛觉得书展有意义，认为（v.）应该去。
ren4wei2/ren2wei3	我刚刚打电话确认了，任伟（n.）应该去。
25.讨厌/陶燕	你不爱运动，讨厌（v.）跑步。
tao3yan4/tan2yan4	为了瘦肚子，陶燕（n.）跑步。
26.主张/主章	刘云重视身体健康，主张（v.）每天按时做饭。
zhu3zhang1	为了让生活有规律，朱章（n.）每天按时做饭。
27.企图/启途	我们早就知道，你企图（v.）制造混乱。
qi3tu2	为了抢劫珠宝，倪启途（n.）制造混乱。
28.提议/题艺	大家刚来到北京的时候，你提议（v.）去爬长城。
ti2yi4	为了感受北京历史文化，李题艺（n.）去爬长城。
29.声明/生鸣	公司出了大纰漏还不及时补救，你声明（v.）这是自掘坟墓。

sheng1ming2	公司出了大纰漏还不及时补救，李生鸣（n.）这是自掘坟墓。
30.担心/丹芯	对附近地区不熟，你担心（v.）可能搭错车。
dan1xin1	对附近地区不熟，李丹芯（n.）可能搭错车。
31.支持/之迟	在辩论比赛中，你支持（v.）回国过春节。
zhi1chi2	为了见到家人，李之迟（n.）回国过春节。
32.生怕/沈坡	刘云不会上网，生怕（v.）买不到票。
sheng1pa4/shen3po1	浏览器打不开，沈坡（n.）买不到票。
33.试图/石涂	李青为了躲避追查，试图（v.）逃到美国。
shi4tu2/shi2tu2	为了躲避仇家追杀，石涂（n.）逃到美国。
34.情愿/晴院	让你磕头认错，你情愿（v.）一头撞死在柱子上。
qing2yuan4	为了表明清白，李晴院（n.）一头撞死在柱子上。
35.禁止/靳芷	这里是公共场合，禁止（v.）吸烟。
jin4zhi3	为摆脱失恋痛苦，靳芷（n.）吸烟。
36.拼命/品茗	为了找对象，你拼命（v.）节食减肥。
pin1ming4/pin3ming2	为了买裙子，倪品茗（n.）节食减肥。
37.允许/云西	这里是吸烟区，允许（v.）抽烟。
yun3xu3/yun2xi1	为了耍帅扮酷，云西（n.）抽烟。
38.准备/准贝	谁都知道，你准备（v.）在班上闹事。
zhun3bei4	报告老师，倪准贝（n.）在班上闹事。
39.继续/姬西	李青不顾城管的阻挠，继续（v.）大声吆喝。

ji4xu4/ji1xi1	想早点卖完红薯回家，姬西（n.）大声吆喝。
40.等于/邓宇	天天玩手机刷微博，等于（v.）在浪费时间。
deng3yu2/deng4yu3	天天玩手机刷微博，邓宇（n.）在浪费时间。
41.忘记/王机	早上出门赶时间，忘记（v.）锁门了。
wang4ji4/wang2ji1	钥匙完全拧不动，王机（n.）锁门了。
42.同意/童艺	刘云为了顾全大局，同意（v.）离开北京。
tong2yi4/ tong2yi4	刘云只有一个要求：童艺（n.）离开北京。
43.以为/易未	你发现教室里没有人，以为（v.）搞错时间了。
yi3wei2/yi4wei4	教室里一个人也没有，易未（n.）搞错时间了。
44.感觉/甘绝	李青听到了后面的脚步声，感觉（v.）被人跟踪了。
gan3jue2/gan1jue2	刘小云从地上的脚印判断：甘绝（n.）被人跟踪了。

8 Appendix 2: name list

牛耀	李硕	庄作	李晴院
舒能	杜友	任伟	靳芷
朱惠	甘雨	陶燕	倪品茗
倪响	陆毅	朱章	云西
尼肯	许耀	倪启途	倪准贝
杜栩	毕虚	李题艺	姬西
杜准	倪山雨	李生鸣	邓宇
李艾	应刚	李丹芯	王机
李贤	柯能	李之迟	童艺
李坡	杜双	沈坡	易未
李叹	李开使	石涂	甘绝