Research Article

Syllable synchronization and the P-center in Cantonese

Ivan Chow⁎, Michel Belyk, Vance Tran, Steven Brown

Department of Psychology, Neuroscience & Behavior, McMaster University, 1280 Main Street West Hamilton, Ont., Canada L8S 4L8

ARTICLE INFO

Article history:
Received 19 November 2013
Received in revised form 1 October 2014
Accepted 16 October 2014

Keywords:
Cantonese
Lexical tone language
P-center
Syllable synchronization
Speech rhythm

ABSTRACT

In speech rhythm analysis, it is important to localize the perceptual center (P-center) of syllables in order to establish a basis for measuring syllabic duration. P-center research has focused primarily on Germanic languages, for which syllables tend to begin with multiple consonants. In Cantonese, the syllable-initial position contains no more than one consonant, making it less prone to durational variation. Studies of various language types using syllables beginning with a singleton indicate that the P-center is generally localized close to vowel onset within the initial consonant–vowel transition. In the present study, Cantonese speakers were found to localize the P-center to the initial consonant–vowel transition when repeating a sequence of identical syllables in a synchronized manner against an audible metronome. However, contrary to previous findings, the metronome beat was aligned close to the onset of the syllable-initial consonant. With increasing repetitions, syllable-initial consonant onset became more closely aligned with the beat while vowel onset became further removed. Similar behavioral patterns are found in sensorimotor synchronization studies (e.g., for finger tapping), suggesting that Cantonese speakers use syllable-initial consonant onset as an articulatory reference point in speech synchronization. Further investigation is needed to elucidate the prosodic and cognitive basis of this behavioral discrepancy.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

To understand speech rhythm, it is essential to examine how timing units align with the syllable (Liberman, Shankweiler, Fischer, & Carter, 1974; Port, 2003). Central to this understanding is the identification of the "perceptual center," or P-center, which is the temporal reference point at which a syllable is perceived to occur (Cooper, Whalen, & Fowler, 1986; Fraisse, 1974; Hoequist, 1983; Marcus, 1981; Morton, Marcus, & Frankish, 1976). It follows that the localization of the P-center within the syllable would provide a basis for measuring syllabic duration (Pompino-Marschall, 1989, 1991).

A paradigm in which a sensorimotor activity, such as finger tapping, is repeated while synchronizing to a visual or auditory metronome signal is commonly used in psychology to study human sensorimotor synchronization (SMS) behavior (see Repp, 2005 and Repp & Su, 2013 for comprehensive reviews). Based on such a paradigm, the present study was designed to observe how Cantonese speakers repeat series of identical monosyllables while synchronizing to an audible metronome click. Without explicitly instructing participants to align the metronome click to a specific acoustic point of reference within the syllable (e.g., syllable-initial consonant onset or vowel onset¹), we attempted to localize the P-center by measuring the timing differences between the temporal reference point (i.e., the time of the metronome click) and the acoustic landmarks used by speakers to mark the syllables' point of occurrence. A similar experimental design was undertaken in P-center studies on Brazilian Portuguese by Barbosa, Arantes, Meireles, and Vieira (2005). Admittedly, this approach is not common in prosody research. As such, it may be subject to some forms of adaptation specific to syllable synchronization that are potentially not used in normal speech. However, if normal speech had “rhythm,” it would be useful to establish a relationship between an isochronous time reference and the point at which a syllable is

1 Abbreviations: P-center, perceptual center; NMA, negative mean asynchrony.
  ⁎ Corresponding author. Tel.: +1 905 525 9140x21443.
  E-mail address: chowiv@mcmaster.ca (I. Chow).
  ¹ In the context of this paper, “onset” is taken to mean the instantaneous point at which a consonant or vowel begins. Onset is not used in the traditional phonological sense to signify the part of a syllable that occurs before the rime.
perceived to occur (i.e., the P-center). The measurement of syllabic duration, then, provides a foundation for performing quantitative analyses on speech rhythm (cf. Allen, 1972; Fraisse, 1982; Klatt, 1976; Lehiste, 1977).

The exact location of the P-center and its corresponding acoustic landmarks continue to be the subject of debate. However, for syllables beginning with a consonant, it is generally agreed that the P-center is localized close to vowel onset, within the transition between the onset of the syllable-initial consonant and the vowel (Barbosa et al., 2005; Hoequist, 1983; Marcus, 1981; Patel, Löfqvist, & Naito, 1999). Research on the P-center is generally conducted from either the perspective of speech perception or speech production. As pointed out by Barbosa et al. (2005), speech perception is linked with speech production in the case of speech synchronization tasks. Thus, the two are either treated as somewhat autonomous systems (Howell, 1988; Pompino-Marschall, 1989; Scott, 1993) or as intimately intertwined (Fowler, 1979, 1983; de Jong, 1992, 1994; Morton et al., 1976). In fact, in order to perform speech synchronization, which involves both processes, it would make sense for both speech perception and production to make reference to the same speech event (Fowler, 1979; de Jong, 1994). This in turn supports the assumption that research from either perspective would yield similar results and that the P-center location would essentially be analogous in both speech perception and production.

In studies that took the speech perception approach (Cooper et al., 1986; de Jong, 1994; Marcus, 1981; Patel et al., 1999; Pompino-Marschall, 1989; Scott, 1998), participants were asked to listen to a pre-recorded isochronous sequence of syllables with different initial consonants and rimes. The methodology of these studies differs mainly in the use of simultaneous or interpolating metronome signals as a temporal reference during the experiment. In the test sequence, syllables were aligned in such a way that the inter-onset intervals (IOI) were constant. Participants were asked to make judgments as to whether the syllable sequence was isochronous or not, either between subsequent syllables or between syllables and the metronome. Based on this judgment, they then adjusted the temporal positions of the syllables until the sequence was perceived as isochronous. The position of the P-center was determined by locating the point in the syllables that coincided with the newly adjusted isochronous interval.

Cooper et al. (1986) carried out a series of speech perception experiments by asking participants to adjust the temporal positions of an isochronous series of syllables with initial fricatives (e.g., /s/) until the series was perceived as isochronous. The duration of these fricatives was varied to create different experimental conditions in order to observe the effects of syllable-initial consonant duration on the P-center location. Variations in the duration of the syllable-initial fricative also triggered changes in the categorical perception of the pre-vocalic consonant in such a way that the fricative at the long end of the spectrum would be perceived as a fricative /s/, while the same fricative would be perceived as an affricate /ts/ at the short end. Results showed that the P-center was not affected by the perceptual (phonetic) categorization of the pre-vocalic consonant. Instead, it was largely influenced by the duration of that consonant and, to a lesser extent, the duration of the syllabic rime, which agrees with findings of Marcus (1981) and Rapp-Holmgren (1971). Using an alternating series of synthetic syllables and metronome signals, Pompino-Marschall (1989) showed similar findings in terms of the syllabic rime. However, in contrast to the work of Cooper et al. (1986) and Hoequist (1983, next paragraph), Pompino-Marschall found a non-linear relationship between the syllable-initial consonant duration and the location of the P-center.

In research that assumed the speech production perspective, participants were asked to produce a sequence of syllables while synchronizing with a visual or auditory metronome signal (e.g., Barbosa et al., 2005; Cummins & Port, 1998, 2009; de Jong, 2001; Hoequist, 1983; Tajima & Port, 2003). De Jong (2001) used the syllable synchronization paradigm to study the temporal relations between open (consonant–vowel) and closed (vowel–consonant) syllables, whereas Cummins and Port’s (1998) research concentrated on observing whether speakers were able to synchronize their speech against external signals that were in phase, anti-phase, or out of phase. The latter did not provide concrete results as to exactly which part of the syllable was used by speakers to align with the metronome signal. Likewise, based on the same “speech cycling” research paradigm (Cummins & Port, 1998, 2009), Tajima and Port (2003) conducted a comparative study to observe how stress patterns of different languages affect the subdivision of timing units between prominent syllables within prosodic feet for English and within words for Japanese. Since the P-center was not the focus of their study, Tajima and Port (2003) assumed that the P-center would be localized to the vowel onset and used it as a reference point for examining the temporal alignment of syllables within a speech cycle. They did not investigate where the P-center was localized, nor did they make a comparison of the P-center location between English and Japanese speakers. Both Hoequist (1983) and Patel et al. (1999) used a paradigm in which participants were asked to repeat alternating monosyllables in an isochronous manner. Hoequist created two test conditions and conducted a comparative experiment between speakers of English, Spanish, and Japanese in which participants were asked to perform the repetition task in phase with a metronome pulse in one condition and without a metronome pulse in the other. However, data collected from the repetition task with a metronome pulse were not included in the statistical analyses. In Patel et al.’s (1999) study, participants were simply asked to produce a sequence of isochronous syllables without the help of a metronome. The P-center was then determined by identifying the location within the syllables that coincided with the point at which isochrony was established. In Barbosa et al. (2005), a speaker of Brazilian Portuguese was asked to repeat 21 series of identical consonant–vowel syllables while synchronizing to metronome signals of four different frequencies (80, 108, 138, and 208 beats per minute (bpm)).

Despite differences in experimental paradigms and the native language of participants, Barbosa et al. (2005), Hoequist (1983), and Patel et al. (1999) generally conclude that the P-center occurs close to vowel onset within the transition between the syllable-initial consonant and the vowel. However, they were unsuccessful in identifying any acoustic or articulatory landmark that precisely aligns with the P-center. On the other hand, Hoequist (1983) found a slight but statistically insignificant language difference in which the alignment of vowel onset to the P-center was less accurate for Spanish speakers compared to that for English and Japanese speakers.Similar to Pompino-Marschall (1989), Barbosa et al. (2005) found the location of the P-center to be largely determined by
the type of the syllable-initial consonant(s). In addition, lower vowels, which produce a higher energy rise in vowel onsets, seem to favor the task of phase synchronization. While synchronization tasks performed at 80, 108, and 138 bpm produced somewhat similar results, phase synchronization was completely lost at 208 bpm. Although Barbosa et al. (2005) claim that the P-center was localized to vowel onsets, they only provided data regarding the time difference between the metronome beat and vowel onsets for different consonant–vowel combinations. Data regarding the timing difference between the metronome beat and the onset of the syllable-initial consonant were missing and thus no comparison was made between the timing of the syllable-initial consonant and vowel onsets. Furthermore, instead of analyzing the reference metronome signal and recorded syllable sequence in a synchronized manner, the P-center was determined by a post-hoc merging of two simultaneous channels: one with the recording of the metronome beat and the other with the recording of the syllables repeated in a synchronized manner against the metronome. Measurement of the time difference between the metronome beat and the syllable began with the first point at which the beat was closest to a syllable. Subsequent P-center locations were then determined in a similar manner. Nonetheless, Barbosa et al. (2005) did not specify which part of the syllable was used as a reference to establish the location of the first P-center. Consequently, subjectivity in locating the P-center severely compromised the validity of their observations. Finally, although the synchronization task was performed in different frequencies with different consonant–vowel combinations, observations were made based on the experimental data collected from a single participant. It is therefore not possible to ascertain whether these observations can be generalized to all speakers of Brazilian Portuguese.

Despite this, research using either approach has yet to yield conclusive results pointing to the acoustic, articulatory, or kinematic landmarks that consistently correspond to the P-center. Particularly relevant critiques about the methodology of past P-center studies come from Villing, Repp, Ward, and Timoney (2011), who indicate that for studies using the perceptual IOI approach and those not using a simultaneous metronome signal as an a priori temporal reference (e.g., Cooper et al., 1986; Patel et al., 1999), it is impossible to empirically derive the absolute P-center locations using only the relative locations implied by the intervals between P-centers without an isochronous, real-time reference. On the other hand, syllable synchronization tasks used in the speech production approach are similar to the experimental design of sensorimotor synchronization (SMS) studies. In both types of studies, participants perform a repeated articulatory activity in an isochronous manner, with or without an external temporal reference (i.e., a visual or auditory metronome). However, syllable synchronization distinguishes itself from other non-verbal SMS tasks in that the reference point of the sensorimotor activity—whether articulatory or acoustic—has yet to be determined, whereas the reference point is made explicit for SMS tasks (e.g., the point at which the fingertip comes into contact with the surface in finger tapping studies).

According to Hoequist (1983), there is evidence suggesting that the reference points on which investigators base their measurements in acoustic speech analysis (i.e., the P-center) should not coincide exactly with those that speakers use to define boundaries of rhythmic units (i.e., the metronome signal). Specifically, rather than looking for an acoustic landmark that lines up with the P-center reference, it may be preferable to identify one that slightly precedes it. This view is supported by Allen (1972) who demonstrated that participants tend to place taps slightly before the onset of periodicity in the acoustic signal when asked to synchronize to an isochronous series of audible monosyllables. This phenomenon, referred to as negative mean asynchrony (NMA), is ubiquitously observed in SMS research that uses an audible metronome signal (see Repp, 2005; Repp & Su, 2013 for comprehensive reviews). However, NMA is not yet fully understood (Aschersleben, 2002; Aschersleben & Prinz, 1995, 1997; Patel, Löfqvist, & Naito, 2005; Repp & Su, 2013). For both musicians and non-musicians alike, NMA is reduced or eliminated with explicit auditory feedback (Aschersleben & Prinz, 1995, 1997; Pressing, 1998). As summarized in Repp and Su (2013), although Hove, Spivey, and Krumhansl (2010) and Repp (2010) did not find any significant effects of musical training on participants’ ability to synchronize to an auditory metronome beat, Repp, London, and Keller (2013) claim that percussionists perform better than other musicians in reducing inter-tap variability (which consists mostly of NMA), while Bailey and Penhune (2010) also show that early-trained musicians perform better than late-trained musicians in synchronization tasks.

As reiterates by Barbosa et al. (2005), most studies on the P-center and speech synchronization have been conducted with speakers of Germanic languages. In these studies, it is generally observed that vowel onset seems to be the main attractor for the P-center—that is, the vowel onset of a syllable tends to be the closest acoustic landmark that participants use as an indication of the point of occurrence of a syllable. The syllable-initial position in Germanic languages commonly occupies a larger proportion of the syllabic duration and is subject to greater variability than in non-Germanic languages. This is illustrated in the words spray, spay and say in English, and stramm (“tight,” pronounced [tram]), Stamm (“tribe, stem or trunk,” [tam]) and Scham (“shame,” [jam]) in German. We speculate that in these languages, the syllable-initial position can consist of up to three consonants, which renders the duration of the pre-vocalic portion highly variable. In contrast, the beginning of the vocalic nucleus of a syllable remains phonologically invariant and perceptually salient, thereby facilitating entrainment to the vowel onset in syllable synchronization tasks for these languages.

With the exception of Xu and Liu’s (2006) study of Mandarin, very little is understood about the P-center phenomenon in East Asian lexical tone languages such as Cantonese, whose phono-temporal structure is markedly different than that of the Germanic languages. Specifically, Cantonese only permits up to one initial consonant and one glide in a syllable, as in 光 ("light," [kwon55]), though most syllables contain only one initial consonant and a small number contain none. Given that the pre-vocalic portion of syllables in Cantonese is shorter and less variable than in Germanic languages, the onset of the syllable-initial consonant may be a more reliable acoustic landmark for the P-center.

Because the syllable represents the basic timing unit in speech (Liberman et al., 1974; Port, 2003), it serves as the most important phonological element in decoding speech rhythm. Cummins and Port (2012) advised against relying on the syllable as an oscillatory system for speech due to idiosyncrasies in pronunciation. For example, in English, “coral” can be spoken as either one or two
syllables. Likewise, in Cantonese syllable fusion occurs in faster speech and affects a small number of high-frequency polysyllabic words in which segments between two syllables are reduced or deleted while maintaining the lexical tones (Wong, 2006). However, at a normal speech rate, syllabic deletion and re-syllabification are rare (Bauer & Benedict, 1997; Perry, Kan, Matthews, & Wong, 2006). Since the syllable is the temporal alignment domain for both segmental and suprasegmental (i.e., tonal) gestures in lexical tone languages (Xu, 1999; Xu & Liu, 2006), the syllable is a more stable timing unit in Cantonese than in non-tonal languages. In addition, in East Asian lexical tone languages, each syllable—which is roughly equivalent to a morpheme or a word—is associated with a particular pitch movement, referred to as a lexical tone. The meaning of a syllable with a given segmental configuration changes when it is associated with a different lexical tone. Table 1 provides a list of the set of lexical tones used in Cantonese, which will be explored in the experiment presented here.

The purpose of the present study was to examine and model the temporal alignment of Cantonese syllables with an external temporal reference provided by a metronome click. This simulated isochronous signal provides a real-time reference for the correspondence between the P-center and the signal. We determined the point at which a syllable in Cantonese coincides with the temporal reference provided by a metronome click. This simulated isochronous signal provides a real-time reference for the will be explored in the experiment presented here. 

Table 1
Cantonese lexical tones. There are six non-checked lexical tones in Cantonese, of which two are rising, three are level, and one is falling. Non-checked tones are associated with open syllables or syllables ending with nasal consonants. Three of the level tones can associate with syllables ending with an unreleased stop to give rise to three additional “checked” tones (Bauer & Benedict, 1997). Tone heights are designated on a 5-point scale (Chao, 1968), in which 1 denotes the lowest relative pitch level and 5 the highest (see Table 1). Contours for the six non-checked lexical tones are therefore transcribed as 55, 25, 33, 23, 21, and 22. The checked tones are represented by a single digit (5, 3, and 2), given their shorter syllabic duration.

<table>
<thead>
<tr>
<th>Lexical tone</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>s55 詩 “poem”</td>
</tr>
<tr>
<td>55 – high level</td>
<td></td>
</tr>
<tr>
<td>25 – high rising</td>
<td></td>
</tr>
<tr>
<td>33 – high-mid level</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>s21 事 “try”</td>
</tr>
<tr>
<td>21 – low falling</td>
<td></td>
</tr>
<tr>
<td>23 – low rising</td>
<td></td>
</tr>
<tr>
<td>22 – low level</td>
<td></td>
</tr>
<tr>
<td>Checked</td>
<td>sik2 色 “to eat”</td>
</tr>
<tr>
<td>5 – high checked</td>
<td></td>
</tr>
<tr>
<td>3 – mid checked</td>
<td></td>
</tr>
<tr>
<td>2 – low checked</td>
<td></td>
</tr>
</tbody>
</table>

2. Methodology

2.1. Participants

Twenty-three native speakers of Standard Cantonese (cf. Gu, Hirose, & Fujisaki, 2006) with literacy in traditional Chinese were recruited for the study (13 males, 10 females; mean age: 19.1; range: 18–21). None of them reported speech or hearing problems. All participants were undergraduate students who received either payment or course credit for their participation. Prior to the start of the experiment, participants provided written informed consent and details on their language and music background. This research was approved by the McMaster University Research Ethics Review Board.

2.2. Syllable repertoire

Participants were asked to repeatedly articulate monosyllables while entraining their speech to an audible, click-like metronome signal. Syllables in the repertoire presented in Appendix 1 were derived from all possible unique combinations of manners of articulation, tones, and rimes, using the long vowel /aː/ in the nucleus position. Compared with other vowels, /aː/ produced the largest number of real words in Cantonese and ensured maximum acoustic contrast with surrounding consonants in subsequent spectrographic analyses. Each manner of articulation was represented by a single initial consonant for consistency (e.g., /ts/ for the affricates). However, if the resulting syllable was not a real word, a different initial consonant with the same manner of articulation was substituted (e.g., for the affricates, [tsa:25] (“bad”) was used in place of [ts’a:25], which is an accidental gap in Cantonese). In addition, checked syllables in combination with the low falling tone do not exist in Cantonese and were therefore excluded from the
repertoire. It should be noted that certain syllables produced with this protocol were not associated with standard dictionary citation tones (Jurafsky, 1988). For example, the word 辣 (“spicy”) was presented with its colloquial pronunciation [laːt̚25] instead of [laːt̚2]. In such situations, there were explicit on-screen instructions to guide the participant to pronounce the tone as desired.

A total of 66 syllables were generated with the above protocol. For the experiment, the syllables were evenly distributed between two groups (A and B) in terms of manner of articulation, tone, and rime. Each participant performed the syllable synchronization task using only one of the two groups of syllables in order to minimize the effects of fatigue and over-practice.

2.3. Procedure

Participants were seated in a sound-attenuating room in front of a computer screen while wearing Koss TD/65 headphones with an Apex 275 headset condenser microphone attachment. In order to create an objective a priori temporal reference for measuring the relative timing of the syllable-initial consonant and vowel onsets (cf. Barbosa et al., 2005; Villing et al., 2011), a 60-bpm audio pulse track was created using Finale (MakeMusic Inc., 2013) and was played through the headphones and simultaneously recorded on a mono-audio track using Adobe Audition 3.0 at a sampling rate of 44.1 kHz. The same track was later superimposed over the recordings while time differences were extracted automatically using Praat (Boersma & Weenink, 2013).

The inter-tap interval limits for finger tapping in SMS studies range from 200 milliseconds to 1800 milliseconds (reported in Repp, 2005: Fraisse, 1982). Although syllabic rate and utterance length are found to vary as a function of a speaker’s age, sex, dialect, occupation, and topic of conversation, research on speech rate indicates that for normal spontaneous speech the overall average speech rate is 196 words per minute, which translates to roughly 3.3 words per second or 300 ms per syllable (cf. Malécot, Johnston, & Kizziar, 1972, for Parisian French spoken in normal conversations; Yuan, Liberman, & Cieri, 2006, for speakers of English and Chinese in conversational telephone speech). In continuous speech, however, syllables are subject to co-articulation. In order to eliminate these effects, we aimed to provide participants with a substantial gap between repetitions with at least the same duration of each syllable. The frequency of syllable repetition therefore needed to be significantly lower than that of spontaneous speech, with inter-pulse intervals of at least 600 ms, or a metronome speed of no greater than 100 bpm. In the initial stage of testing, three metronome speeds were created from 60 to 100 bpm (20 bpm apart), and through four pilot tests the 60-bpm track was found to be the most comfortable speed for the repetition task.

Ideally, an individualized metronome track would have been created at the beginning of each trial to match the natural syllable production rate of each participant. However, due to equipment limitations, a metronome track with the duration of the experimental session (21 min) could not be generated spontaneously but rather needed to be created through a real-time recording. This could not be achieved on the spot in the experimental timeslot that participants were allotted; therefore, the metronome track was pre-recorded at a fixed frequency. The 21-min metronome track was played continuously on the headphones throughout the experimental session. While it was not feasible to create individualized metronome tracks, the use of a fixed frequency for all participants is justified given the aforementioned speech rate parameters and pilot testing results. Moreover, we confirmed that after participants had practiced the synchronization task at 60 bpm in the warm-up stage of the experiment, none experienced any issues with the specified frequency during the actual trial.

Participants’ vocal output was recorded through the microphone onto a separate audio track using the same parameters. The experimenter remained out of view at the back of the room for the duration of the experiment in order to ensure adherence to the protocol. All procedural information and experimental stimuli were displayed on a computer screen using Presentation (Neurobehavioral Systems, 2013) (see Fig. 1).

For each trial, a primer screen showed the target syllable in the context of a two- or three-syllable compound word for 3 s in order to prompt the desired pronunciation of the target syllable. For example, when displayed in the compound word 差人 (“policeman,” [tsʰa:j55jan21]), the character 差 has two pronunciations: [tsʰa:55] (“bad, inferior”) and [tsʰa:j55] (“work-related errands”). In this context, the character was pronounced [tsʰa:j55]. Each context word was annotated with tone numbers, an English translation, and
Romanized transcriptions and tone numbers (Chao, 1968) with which Cantonese speakers are relatively more familiar than IPA. The target syllable contained within the context word was bolded and underlined. A complete list of the context words is provided in Appendix 2. Next, a “Ready” screen initiated a 3-s countdown, after which the participant was asked to articulate the target syllable successively in sync with the beat of the metronome. At the end of a trial, a “Rest” screen was displayed, followed by a “Here comes the next syllable” screen before the next trial began. Each experiment consisted of two practice trials and 33 experimental trials, with syllables presented in a random sequence from either stimulus group A or B. During the initial briefing, participants were instructed to practice the task by repeating the target syllable 10 times in each trial. However, 13 repetitions were collected for most trials because each trial lasted 12 s and participants were not required to keep track of their number of repetitions during the actual trials.

2.4. Data analysis

Data analysis was carried out at the acoustic level using Praat (Boersma & Weenink, 2013). Recorded data from two separate tracks (i.e., the metronome and the participant’s voice) were synchronized and analyzed across three tiers. On the first tier, an in-house Praat script was used to insert a boundary at the highest point of each intensity spike triggered by the metronome signal. The remaining two tiers were used for speech analysis. On the second tier, a boundary was manually placed at the beginning of each target syllable’s initial consonant. As mentioned above, 13 syllable repetitions were collected for most trials; the remaining two tiers were used for speech analysis. On the second tier, a boundary was manually placed at the beginning of each target syllable’s initial consonant. As mentioned above, 13 syllable repetitions were collected for most trials; the first and last were discarded due to transient effects (Klatt, 1976). Table 2 shows the acoustic landmarks used to identify the syllable-initial consonants for the different manners of articulation. Syllables beginning with a vowel, such as “late,” [a:n33], are occasionally pronounced allophonically (i.e., with a syllable-initial consonant) as either [ʔa:n33] (glottal stop) or [ʔa:n33] (velar nasal). These allophones were categorized as different manners of articulation: those with /l/ formed their own category whereas those beginning with /n/ were analyzed as nasals. On the third tier, boundaries were manually placed where the vowel is articulated (i.e., the vowel onset), as indicated by the start of formant patterns for /a:/ in the spectrogram. The placement of all boundaries was based on waveform and spectrographic indices.

The time difference between the metronome signal and the syllable-initial consonant or vowel onsets for each target syllable was extracted using an in-house Praat script. All statistical data analyses, including ANOVAs, regressions, and t-tests, were performed using SPSS (IBM Software, 2013). R (R-Project, 2013) was used exclusively for producing figures.

3. Results

Fig. 2 shows the timing of syllable-initial consonant and vowel onsets relative to the metronome beat for various manners of articulation. Regardless of whether a syllable begins with a consonant or vowel, the onset of the syllable was closely aligned with the metronome beat. For syllables with an initial consonant, vowel onset followed the metronome beat more distantly. On average, syllable-initial consonant onset preceded the time of the metronome beat by 15.77 ms (SD = 37.10 ms). However, a one-sample t-test indicated that the time difference between the two events was below the level of significance (t (22) = −2.038, p = 0.054), whereas vowel onset occurred on average 74.62 ms (SD = 36.75 ms) after the metronome beat (t (22) = 9.738, p < 0.05). Overall, syllable-initial consonant onset occurred 0.4 standard deviations before the metronome beat, while vowel onset occurred 2.0 standard deviations after it. A series of one-way ANOVAs showed group (between-subject) differences to be significant in both the timing of syllable-initial consonant onsets (F (22, 766) = 12.482, p < 0.05) and vowel onsets (F (22, 766) = 13.801, p < 0.05).

In Fig. 2, it is observed that the manner of articulation strongly influenced participants’ timing with the metronome. For syllables beginning either with a vowel or a consonant of short segmental duration (e.g., stops or glottal stops), syllable onset occurred shortly after the metronome beat. For syllable-initial consonants with inherently long segmental durations (e.g., fricatives, nasals, laterals, and affricates), participants began their articulatory gestures before the metronome beat. Fricatives with the longest segmental durations were an exceptional case in that syllable-initial consonant onset was placed farther from the metronome beat than was vowel onset. For syllables beginning with a vowel (i.e., without an initial consonant), vowel onset closely followed the metronome beat. For syllables that were pronounced with initial allophonic consonants /l/ and /n/, alignment with the metronome was similar to syllables beginning with a consonant of the corresponding manner of articulation. As mentioned in the Methods section, syllables beginning with /l/ are shown under the “glottal” category in Fig. 2 and those beginning with /n/ are shown under the “nasal” category.
A one-way ANOVA revealed that the manner of articulation of the initial consonant had a significant influence on the timing of initial-consonant and vowel onsets. However, when comparing affricates, the only set of consonants included in our dataset was for aspirated and non-aspirated counterparts (/ts/ vs. /tsʰ/). Results from a two-tailed T-test showed that aspiration had no significant influence on the timing of initial consonant or vowel onset (Table 3). Regressions were performed to examine the relationship between the duration of the syllable-initial consonant and the timing of both syllable-initial consonant and vowel onsets relative to the metronome beat. While the correlation between the duration of the syllable-initial consonant and the timing of its onset failed to reach significance ($r^2 = 0.154$, $F(1, 21) = 3.809$, $p = 0.064$), the correlation between the duration of the syllable-initial consonant and the timing of vowel onset was far weaker ($r^2 = 0.002$, $F(1, 21) = 0.43$, $p = 0.84$).

Next, we examined timing as a function of syllable repetition. Fig. 3 shows the overall timing of syllable-initial consonant and vowel onsets relative to the metronome beat as a function of the number of repetitions of test syllables in a given trial. Syllables beginning with a vowel were excluded. As can be seen, syllable-initial consonant onsets gradually converged with the timing of the metronome beat whereas vowel onsets diverged from it. The timing of syllable-initial consonant and vowel onsets progressed in a non-linear fashion, which indicates that participants were not simply locked into a phase with a fixed frequency lower than that of the metronome. Rather, Fig. 3 suggests a pattern of self-correction in which the time difference between syllable-initial consonant onsets and the metronome beat was gradually reduced as the repetition task progressed. This is further supported by the progressive decrease of standard deviations for the time differences between the metronome beat and syllable-initial consonant onsets as the number of repetitions increased. A linear regression showed that the position within the repetition sequences strongly and negatively correlated with the standard deviations ($r^2 = 0.964$, $F(1, 9) = 243.527$, $p < 0.01$).

With regard to participants’ musical background, data was collected on the number of years of musical training, which was then divided into four categories: none, 0–5, 6–10, >10. ANOVA results demonstrated that musical background had no significant influence on the timing of initial-consonant onsets and the metronome ($F(3, 19) = 2.801$, $p = 0.068$). Finally, we analyzed the influence of rime type, tone identity, tone type, and tone height on the timing of syllable-initial consonant and vowel onsets. Table 4 presents a summary of the factors and their statistical results using a series of one-way ANOVAs. None of these variables had a significant influence on the timing of syllable-initial consonant or vowel onsets. Since the main effects for rime type and tone were insignificant, it was deemed unnecessary to test the significance of their interaction.
4. Discussion

4.1. Syllable vs. vowel onsets

In order to develop a better understanding of Cantonese speech rhythm, we aimed to localize the P-center in Cantonese using a syllable synchronization experiment. Our results agree with previous conclusions that the P-center falls within the initial consonant–vowel transition of syllables. However, we additionally remarked that while neither was aligned perfectly with the metronome beat, the onset of the syllable-initial consonant was aligned more closely than the onset of the vowel. An exception to this observation is for syllables starting with a vowel, in which vowel onset was found to closely align with the metronome. In both cases, it is the very start of the syllable (which we term the “syllable onset”) that most closely synchronized with the metronome. This finding suggests that the P-center in Cantonese is mapped to syllable onset, which differs from trends seen with Germanic languages (Cooper et al., 1986; de Jong, 1994; Marcus, 1981; Morton et al., 1976), Brazilian Portuguese (Barbosa et al., 2005), and Spanish (Hoequist, 1983) and Japanese (Hoequist, 1983; Tajima & Port, 2003). More specifically, the present results contrast with studies that have used syllables beginning with a singleton in both Germanic and non-Germanic languages (Barbosa et al., 2005; Hoequist, 1983; Pompino-Marschall, 1989). Furthermore, Table 3 illustrates that for Cantonese, the manner of articulation of syllable-initial consonants was a significant factor in determining the location of the P-center. Results from regression tests did not show any linear relationship between syllable-initial consonant duration and the timing of both syllable-initial consonant and vowel onsets, although the correlation between syllable-initial consonant duration and timing was shown to be marginally insignificant. These results are contradictory to observations of Cooper et al. (1986), Hoequist (1983), Marcus (1981), Rapp-Holmgren (1971), but consistent with those of Pompino-Marschall (1989). The comparison between aspirated and non-aspirated counterparts of affricates showed that aspiration had no significant influence on the location of the P-center. This was expected since syllable-initial consonant duration is largely affected by the manner of articulation and to a much lesser extent the presence of aspiration.

The claim that the P-center is localized to vowel onset largely originated from studies adopting the speech perception approach without use of a metronome signal as a real-time reference (Cooper et al., 1986; Marcus, 1981; Patel et al., 1999; Scott, 1998, etc.). Given that the speech production studies of de Jong (2001), Cummins and Port (1998), and Tajima and Port (2003) did not focus on locating the P-center, the only studies that took a similar approach to ours are those of Hoequist (1983) and Barbosa et al. (2005).

As mentioned in the Section 1, in Hoequist’s study, participants were asked to repeat series of monosyllables with and without synchronizing to a metronome beat. However, data collected from the repetition task with a metronome pulse were not included in the statistical analyses. In Barbosa et al. (2005), participants repeated series of monosyllables beginning with a singleton while synchronizing to a metronome of different frequencies that were interpolated onto the recorded monosyllables during speech analysis. Alignment of the metronome signal was based on the researchers’ own perception of what was temporally “closest” to the
syllable, a location which they failed to specify in their report. The subjectivity of such speech perception-based studies has been a main critique of Villing et al. (2011). The lack of a real-time reference raises questions about the validity of Barbosa et al.'s (2005) and Hoequist's (1983) claims that the P-center is localized to vowel onset. Avoiding this pitfall (cf. Villing et al., 2011), the present study attempted to locate the P-center by comparing the time differences between syllable or vowel onsets and an explicit, objective temporal reference point provided by the beat of the metronome. As such, it is possible that the differences in syllable synchronization behavior observed between past studies and the present one may be due to differences in experimental design and paradigm rather than inherent behavioral differences between Cantonese speakers and those of previously examined languages. Regardless, the incongruence of the present results with previous research warrants further investigation. A conceivable first step would be to replicate the present experiment with speakers of other languages in order to determine whether P-center location is affected by native language experience.

4.2. Negative mean asynchrony

In finger tapping studies, when participants are instructed to align their taps with a metronome, they tend to tap before rather than on the beat (Aschersleben, 2002; Haken et al., 1985; Repp, 2005), an anticipatory phenomenon referred to as negative mean asynchrony (NMA). In the present study, NMA was observed for syllable onsets but not for vowel onsets. Although finger tapping and the pronunciation of monosyllables have different neural bases, similar durational patterns were observed in both activities. With increasing repetitions of a syllable, syllable onset became more closely aligned with the metronome while vowel onset became further removed. This improvement in the accuracy of syllable-initial consonant alignment is consistent with research on NMA. In finger tapping studies, NMA decreases over time with explicit auditory feedback (Aschersleben & Prinz, 1995, 1997; Pressing, 1998).

In the present experiment, explicit auditory feedback provided by the metronome decreased NMA, but only for syllable onsets. The opposite trend was observed for vowel onsets. This lends support to our hypothesis that Cantonese speakers perceive syllable onset as the start of the syllable, and hence improve syllable-initial consonant alignment with the metronome through repetition.

Agreeing with the results of Repp (2010) and Hove et al. (2010), we found that musical background was an insignificant factor in the alignment between the metronome beat and syllable-initial consonant onset. Moreover, since none of the participants in our experiment identified themselves as percussionists, it was not possible to verify Repp et al. (2013) claim. Although NMA has yet to be explored in the context of syllable synchronization, its presence in the present experiment provides an additional clue about the location of the P-center in Cantonese.

4.3. Other independent factors

No factors related to rime and lexical tone had an effect on the location of the P-center. The non-significant effect of rime type is consistent with previous research demonstrating that the location of the P-center is predominantly determined by the duration of syllable-initial consonants (Cooper et al., 1986; Marcus, 1981; Morton et al., 1976). As previously mentioned, however, the location of the P-center varied with the syllable-initial consonant's manner of articulation. Duration varied systematically among different manners of articulation, and this variation was directly related to the temporal location of the P-center. Finally, the fact that neither the identity, type, nor height of a tone affected the P-center location supports the view that differences in syllable synchronization behavior between speakers of Cantonese and the Germanic languages are the result of segmental (or phonotactic) differences in the syllable-initial position and not suprasegmental parameters related to lexical tone such as syllable pitch.

4.4. Limitations

In the syllable repertoire, checked syllables associated with the high tone (5) corresponded to either obscure words or non-words in Cantonese and were therefore excluded from the repertoire. However, this had only a negligible impact on the results of the present experiment, since we demonstrated that tone identity, type, and height did not have a significant effect on the location of the P-center.

Finally, as indicated by the ANOVA results, there were significant group (between-subject) differences in the timing of syllable-initial consonant and vowel onsets relative to the metronome beat. Pompino-Marschall (1989) similarly found significant between-subject differences in his syllable synchronization experiment due to idiosyncratic speech patterns exhibited by one particular participant. Although our results showed a tendency for Cantonese speakers to align syllable-initial consonant onset with the metronome beat in a syllable synchronization task, participants behaved quite differently from one another. In order to better understand this tendency, it would be worthwhile to replicate the present experiment with a larger group of participants.

5. Conclusion

The present study demonstrated that, in contrast to past syllable synchronization work with various types of languages, speakers of Cantonese use the onset of syllables rather than the onset of vowels to align with a metronome. We speculate that this is due to the fact that syllable-initial consonants are shorter and less variable in East Asian languages than in Germanic languages. It is this decreased duration and variability that allow syllable onsets to be more reliable acoustic landmarks for the P-center in Cantonese. However, because of the differences in methodology between previous studies and the present one, it is difficult to verify whether
native language experience is an important factor in syllable alignment behavior. Nonetheless, unlike many of the past P-center investigations (e.g., Barbosa et al., 2005; Cooper et al., 1986; Hoequist, 1983), the current study based the relative timing of syllable-initial consonant and vowel onsets on an objective temporal reference in both the experiment and analysis, which in turn strengthens the validity of our findings.

Finally, in comparing the metronome alignment patterns of the syllable-initial consonant and vowel onsets across a series of test syllables, the effect of negative mean asynchrony (NMA) was only observed for syllable-initial consonants. In other words, participants exhibited improved accuracy in their ability to align syllable-initial consonant onset but not vowel onset as a function of repetition, supporting our hypothesis that syllable onset is indeed used as the P-center in Cantonese. Identification of the P-center enables us to more accurately measure syllabic duration and better understand speech rhythm in Cantonese. In future research, we hope to adapt this experimental design to study the speech rhythm of other languages in order to determine whether native language experience is related to the discrepancy in P-center alignment behavior between the present and past studies.

Acknowledgments

This work was funded by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) awarded to S.B (grant number: 371336). We are grateful to Matthew Poon for his contribution to the pilot experiments, data collection, and analysis, and to Alan Kennedy for his help with fine-tuning the design of the figures and tables.

Appendix 1

List of target syllables. * = Character was not pronounced as per standard dictionary citation tones. Instructions were provided to prompt the desired pronunciation. ∅ = Not applicable. Syllables belonging to Group A are displayed in regular type font, while those in Group B are displayed in italicized font.

<table>
<thead>
<tr>
<th>Rime</th>
<th>Level Tone (55, 33, 22)</th>
<th>Falling Tone (21)</th>
<th>Rising Tone (25, 23)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop</td>
<td>Ø</td>
<td>Diphthong</td>
<td>Nasal</td>
</tr>
<tr>
<td>Nas</td>
<td>媽</td>
<td>ma:55</td>
<td>麻</td>
</tr>
</tbody>
</table>

Appendix 2

List of context words for the target syllables. * = Character was not pronounced as per standard dictionary citation tones. Instructions were provided to prompt the desired pronunciation. ∅ = Not applicable. Syllables belonging to Group A are displayed in regular type font, while those in Group B are displayed in italicized font.


