Optimal Syllabic Rates and Processing Units in Perceiving Mandarin Spoken Sentences

Guangting Mai1, Gang Peng1,2

1 Language Engineering Laboratory, The Chinese University of Hong Kong, Hong Kong
2 Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
gtmai@ee.cuhk.edu.hk, gpeng@ee.cuhk.edu.hk

Abstract

This paper presents our investigations on the syllable-related processing during human perception of Mandarin spoken sentences. Two behavioral perception experiments were conducted employing a signal synthesis method in a previous study [1]. We found (1) a clear relationship between speech intelligibility and syllabic rates of spoken sentences and (2) significantly higher speech intelligibility of sentences acoustically segmented at sub-syllable and syllable levels than at the level beyond one syllable. We therefore revealed the optimal syllabic rates and processing units in perceiving Mandarin continuous speech and further discussed the association between our results and the possible underlying neural mechanisms in the human brain.

Index Terms: Mandarin, optimal syllabic rates, optimal processing units, speech intelligibility

1. Introduction

Syllable is an important phonetic unit in daily speech communications, and speech sounds are normally transmitted at relatively steady and restricted range of syllabic rates at 4–8 Hz across different languages according to various production data [2]. This speech rhythm phenomenon and the syllable-related processes of how human beings integrate speech sounds into linguistically meaningful percepts are fundamental topics for speech science.

From the neurophysiological aspect, there are several studies in recent years which tried to demonstrate that such syllable-related processing does exist and should originate from the neural mechanisms within the human brain. Poeppel proposed a hypothesis [3] which suggests there are two different time scales of ‘integration windows’ associated with the neural oscillations in the auditory cortices (the short-scale windows range approximately from 20–40 ms while the long-scale windows range from 150–250 ms). The brain processes and integrates different linguistic information through such ‘windows’. Zatorre et al. [4] also suggested a similar model stating that the short-scale windows would extract information with relatively fast changing modulations (e.g., consonant-vowel transitions), while the long-scale windows would extract information with relatively slow changing modulations such as syllables, long vowels and lexical tones. Accordingly, these hypothesized ‘integration windows’ would be likely to play a specific role in processing speech signals with perceptual units at syllable or sub-syllable levels.

Real neurophysiological data were also provided to support the existence of such ‘integration windows’. Recently, Giraud et al. [5] tested the endogenous neural oscillations in human auditory cortices and motor regions. They found there are 3–9 Hz neural oscillations and 28–40 Hz oscillations (corresponding to the brain ‘integration windows’ of around 100–300 ms and around 25–40 ms, respectively, which are quite similar with the ranges proposed by Poeppel [3]) in the motor regions responsible for mouth and tongue movements and the auditory cortices. In this study, the oscillations at 3–9 Hz basically match the general range of the syllabic rates commonly produced and therefore were hypothesized to be related to the syllable level processing in speech perception.

Although these studies have provided neurophysiological hypothesis or evidence on the existence of the syllable or sub-syllable level processing, there is still no evidence that supports the hypothesis that the optimal syllabic rates for Mandarin speech perception are aligned with the normal range in speech production (4–8 Hz); (2) it is not clear whether there actually exists such speech analysis processes and ‘integration time windows’ at syllable and sub-syllable levels in human speech perception.

We conducted two behavioral experiments on Mandarin speech perception employing a signal synthesis method of ‘Time-Compression and Silence-Insertion’ modified from the recent work by Ghitza and Greenberg [1] and further tested the above two questions. In Experiment 1 (Exp. 1), we employed this method and obtained a positive answer for the first question that decline in intelligibility can be caused by the disruption in the syllabic rates beyond those at around 3–8 Hz (which is quite consistent with the normal range of production data in Mandarin [2]). In Experiment 2 (Exp. 2), we compared the intelligibility of spoken sentences which were acoustically segmented at sub-syllable, syllable levels and the level beyond one syllable. We then found the optimal processing units when perceiving spoken sentences and discussed the possible relation between such units and the syllable and sub-syllable level processing in Mandarin speech perception.

2. Experiments and Results

2.1. Subjects and tasks

300 different Mandarin Semantically Unpredictable Sentences (SUSs, 140 for Exp. 1 and 160 for Exp. 2) were naturally produced at 4-Hz syllabic rate by a single male native Mandarin speaker sampled at 44.1 kHz for the signal processing for both experiments (details in Section 2.2 and 2.3). The SUSs are sentences that are syntactically acceptable but semantically anomalous [8]. Each Mandarin SUS was constituted of several disyllabic words like “地图突破严肃的地点” (which means ‘maps break through the serious place’). The objective of using SUSs was to reduce the guessing effect from contextual information. All the words in SUSs were chosen manually from the 9,000 most frequent items of a Mandarin word frequency corpus compiled by the Institute of Linguistics of Academia Sinica in Taiwan [9]. 13 (9 males and 4 females, mean age = 22.4 yr (S.D. = 0.9)) and 15 (10 males...
and 5 females, mean age = 22.5 yr (S.D. = 0.8)) native Mandarin speakers at Sun Yat-sen University in Canton, P.R. China were recruited for Exp. 1 and Exp. 2, respectively. No participants reported any hearing and speech difficulties or neurological pathology. During the experiments, they were instructed to listen to the SUSs and write down the words or syllables they heard. The volume was adjusted to a comfortable level during the training session and each SUS played only once during the formal test sessions.

2.2. Experiment 1

2.2.1. Stimuli Preparation

Figure 1: Examples for describing the preparation procedure of a sample sentence ‘工作庆祝广泛的特征’ (‘Jobs celebrate the vast traits’). (A) Originally recorded speech at 4-Hz syllabic rate. (B) Time-compressed speech with 14.7-Hz syllabic rate, i.e. 68 ms per syllable. (C) Speech of (1/2)SYL condition (i.e., condition with 34-ms speech fragment) with periodically inserted 46-ms silence. (D) An ‘Acoustic Unit’ (a speech fragment plus an inserted silence interval)

The signal processing procedures of preparing the SUS stimuli in Exp. 1 are modified from the synthesis method of ‘Time-Compression and Silence-Insertion’ in recent study by Ghita and Greenberg [1] and described as below.

Firstly, the waveforms of all sentences were initially recorded at 4-Hz syllabic rate as mentioned in Section 2.1 and then time-compressed to the syllabic rate of 14.7 Hz through the PSOLA algorithm in PRAAT [10], where the average syllable length is about 68 ms. The intelligibility of such time-compressed waveforms is poor (syllable errors of 47%, as shown in Table 1).

Secondly, the compressed waveforms were then equally segmented into successive speech fragments. We then created two conditions: (1) the condition with each speech fragment length of 34 ms that equals to the average duration of half a syllable in the compressed waveforms (we thus designate this condition as the ‘(1/2)SYL’ condition); (2) the condition with each speech fragment length of 68 ms that equals to one syllable duration in average (the ‘1SYL’ condition). The aim of creating such two conditions is that we’d like to see if different performances can be observed between such speech signals segmented at sub-syllable level and syllable level.

Thirdly, silent intervals were periodically inserted after each speech fragment using Matlab (see Fig. 1 for example). Table 1 shows all 14 cases created by this procedure. An ‘Acoustic Unit’ is defined as a unit that includes one speech fragment plus its following adjacent silence interval. Accordingly, in the (1/2)SYL condition, ‘syllable length’ in Table 1 refers to the average syllable duration in the synthesized waveforms and equals to the length of two successive Acoustic Units, each of which contains half a syllable information; in the 1SYL condition, the syllable length equals to the length of one Acoustic Unit.

Through this procedure, we created sentences with the same speech contents as in the compressed speech signals but with different syllabic rates ranging from 2 to 14.7 Hz (see Table 1). Therefore, we were able to explore how speech intelligibility is modulated by sentences’ syllabic rate by calculating the speech identification accuracies. Furthermore, following the manipulations in [1], boundaries at both sides of each speech fragment were multiplied by a 1-ms-long (rise/fall time) cosine-shaped window to reduce the transient effect. In order to perceptually mask the speech discontinuities caused by silence, a speech-shaped noise was added to each signal after silence insertions. The noise level was adjusted to -30 dB relative to the power of the signal prior to insertions.

There were 140 different synthesized SUSs in Exp. 1 and they were interleaved during the formal test sessions.

<table>
<thead>
<tr>
<th>Speech fragment length</th>
<th>Silence length (ms)</th>
<th>Acoustic Unit (ms)</th>
<th>Syllable length (ms)</th>
<th>Syllabic rate (Hz)</th>
<th>Syllable error %</th>
</tr>
</thead>
<tbody>
<tr>
<td>34 or 68 ms</td>
<td>0</td>
<td>34 or 68</td>
<td>68</td>
<td>14.7</td>
<td>47.0</td>
</tr>
<tr>
<td>(1/2)SYL (34 ms)</td>
<td>26</td>
<td>60</td>
<td>120</td>
<td>8</td>
<td>23.8</td>
</tr>
<tr>
<td></td>
<td>46</td>
<td>80</td>
<td>160</td>
<td>6</td>
<td>18.2</td>
</tr>
<tr>
<td></td>
<td>86</td>
<td>120</td>
<td>240</td>
<td>4</td>
<td>20.4</td>
</tr>
<tr>
<td></td>
<td>126</td>
<td>160</td>
<td>320</td>
<td>3</td>
<td>23.1</td>
</tr>
<tr>
<td></td>
<td>166</td>
<td>200</td>
<td>400</td>
<td>2.5</td>
<td>29.1</td>
</tr>
<tr>
<td></td>
<td>216</td>
<td>250</td>
<td>500</td>
<td>2.5</td>
<td>38.4</td>
</tr>
<tr>
<td>1SYL (68 ms)</td>
<td>12</td>
<td>80</td>
<td>80</td>
<td>12</td>
<td>35.1</td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>120</td>
<td>120</td>
<td>8</td>
<td>22.2</td>
</tr>
<tr>
<td></td>
<td>92</td>
<td>160</td>
<td>160</td>
<td>6</td>
<td>22.4</td>
</tr>
<tr>
<td></td>
<td>132</td>
<td>200</td>
<td>200</td>
<td>5</td>
<td>22.2</td>
</tr>
<tr>
<td></td>
<td>182</td>
<td>250</td>
<td>250</td>
<td>4</td>
<td>29.0</td>
</tr>
<tr>
<td></td>
<td>262</td>
<td>330</td>
<td>330</td>
<td>3</td>
<td>35.6</td>
</tr>
<tr>
<td></td>
<td>332</td>
<td>400</td>
<td>400</td>
<td>2.5</td>
<td>38.9</td>
</tr>
</tbody>
</table>

Table 1. All the 14 cases in Exp. 1

Figure 2: (A) Syllable error rate of the (1/2)SYL and 1SYL conditions as a function of syllable length. (B) Comparison of syllable error rates between the (1/2)SYL and 1SYL conditions. Error bars represent the SEM across all the 13 subjects. ***, p<0.01

Fig. 2(A) shows the syllable error rates for the (1/2)SYL and 1SYL condition averaged over the 13 subjects. The X axis represents the average syllable length, which inversely correlates with the syllabic rate (numerical data in Table 1).

U-shaped curves were obtained for both conditions, which are similar to the results in Ghita and Greenberg’s work [1] with periodical insertions of silence. For the (1/2)SYL condition, the error rate is the lowest when the syllabic length is 160 ms (at around 6-Hz syllabic rate; define as ‘×160 case’) and no significant difference was found between ×120, ×160, ×240 and ×320 cases (at syllabic rates of around 3–8 Hz; in all pairwise comparisons, p > 0.11). For 1SYL condition, the error rate reaches the lowest when the syllable length is 120, 160 and 200 ms (×120, ×160 and ×200 cases, at syllabic rates of 5–8 Hz). The error rates are no more than 25% at all the cases mentioned above.

On the other hand, however, both at fairly short and long syllable length cases, there exists sharp increases in syllable
identification errors. For example, the error rate of the far left point (which represents the time-compressed SUSs with syllable length of 68 ms (at the syllabic rate of 14.7 Hz)) is close to 50%. And at long syllable length cases such as >500 case in (1/2)SYL condition and >400 case in 1SYL condition, the error rates are up to around 40%. It thus reveals that the Mandarin SUS intelligibility is modulated by syllabic rates and the optimal range of the syllabic rates should be 3–8 Hz.

Furthermore, it is noteworthy that subjects generally have better performances in the (1/2)SYL condition than the 1SYL condition (Fig 2(A)), especially at low syllabic rates. Fig 2(B) illustrates the syllable error rates comparing these two conditions: the gray and white bars represent the errors of the (1/2)SYL (errors averaged across ×120, ×160, ×240, ×320 and ×400 cases) and the 1SYL condition (errors averaged across ×120, ×160, ×200, ×250, ×330 and ×400 cases), respectively. The cases for this comparison were chosen according to the range of syllabic rates shared by these two conditions (syllable lengths of 120–400 ms, see Fig 2(A)). Statistical analysis showed that errors of the (1/2)SYL are significantly lower than the 1SYL condition (p < 0.01). This result indicates that subjects may preferentially recognize Mandarin speech signals in a way where the speech information is more evenly distributed and treat half a syllable as a basic perceptual unit ((1/2)SYL) rather than one syllable (1SYL).

### 2.3. Experiment 2

#### 2.3.1. Stimuli Preparation

In order to compare the syllable identification performances of SUSs segmented at sub-syllable, syllable levels and the level beyond one syllable, Exp. 2 was conducted. The stimuli preparation procedures were similar to Exp. 1, and the main difference was that the syllable rates of the sentences were kept constant to 4 Hz and 6 Hz, respectively, which are both in the optimal range of syllabic rates for syllable intelligibility obtained in Exp. 1. Fig. 3(A) is the original speech at the 4-Hz syllabic rate. The waveform was firstly time-compressed to the syllabic rate of 14.7 Hz. The compressed waveform was then equally segmented into successive speech fragments (22 ms in Fig. 3(B) and 89 ms in Fig. 3(C)). Finally silences with proper lengths (58 ms in Fig. 3(B) and 241 ms in Fig. 3(C)) were inserted so that the sentence length went back to approximately the same as the original sentence. Meanwhile the syllabic rate went back to around 4 Hz. Table 2 shows all 16 cases in Exp. 2.

The most important parameter is the number of syllables within each Acoustic Unit. We define the ‘≤1SYL’ condition as the condition where each Acoustic Unit contains no more than one syllable (namely, speech fragment length ≤ 68 ms, labeled as single asterisks in Table 2), and the ‘>1SYL’ condition as the condition where each Acoustic Unit contains more than one syllable (namely, speech fragment length > 68 ms, labeled as double asterisks in Table 2).

There were totally 160 different SUSs in Exp. 2.

<table>
<thead>
<tr>
<th>syllabic rate (Hz)</th>
<th>speech fragment (ms)</th>
<th>silence length (ms)</th>
<th>Acoustic Unit (ms)</th>
<th>syllable numbers in each AU†</th>
<th>syllable error rate%</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Hz (around 6 Hz)</td>
<td>11</td>
<td>29</td>
<td>40</td>
<td>0.16*</td>
<td>21.3</td>
</tr>
<tr>
<td>250 ms per syllable</td>
<td>27</td>
<td>73</td>
<td>100</td>
<td>0.4*</td>
<td>15.7</td>
</tr>
<tr>
<td>119</td>
<td>43</td>
<td>117</td>
<td>160</td>
<td>0.64*</td>
<td>25.6</td>
</tr>
<tr>
<td>68</td>
<td>54</td>
<td>146</td>
<td>200</td>
<td>0.8*</td>
<td>31.1</td>
</tr>
<tr>
<td>89</td>
<td>68</td>
<td>182</td>
<td>250</td>
<td>1*</td>
<td>28.8</td>
</tr>
<tr>
<td>241</td>
<td>89</td>
<td>241</td>
<td>330</td>
<td>1.32**</td>
<td>47.2</td>
</tr>
</tbody>
</table>

#### 2.3.2. Results

Syllable error rate as a function of number of syllables within each Acoustic Unit are illustrated as Fig. 4(A). Detailed data are shown in Table 2. In Fig. 4(A), the half a syllable and one syllable boundaries are labeled (the vertical dashed lines).

Consistent with the results in Exp. 1, the cases that each Acoustic Unit contains around half a syllable (corresponds to the (1/2)SYL condition in Exp. 1) have lower errors than the cases that each Acoustic Unit contains around one syllable (corresponds to the ‘1SYL’ condition): (1) for 4-Hz syllabic rate SUSs, the error of the ×0.48 case (0.48 syllable within each Acoustic Unit) is significantly lower than the ×1 case (1 syllable within each Acoustic Unit) (p < 0.05); (2) for 6-Hz syllabic rate SUSs, the error of the ×0.48 case is lower than the ×0.96 case by 3% (detailed numerical data in Table 2), though no significant difference was found between them (p = 0.28).

What we are particularly concerned about is whether speech could still be identified efficiently at the level beyond one syllable, so we compared the >1SYL (acoustically segmented at the level beyond one syllable) with the ≤1SYL condition (acoustically segmented at sub-syllable and syllable level). It can be seen from Fig. 4(A) that sharp increases in syllable errors happen when each Acoustic Unit contains more than one syllable information (>1SYL) compared to the cases...
in the ≤1SYL condition, both for 4- and 6-Hz syllabic rate SUSs. Although there is a sharp drop at the ×1.98 case in the >1SYL condition (in which each Acoustic Unit contains a disyllabic Chinese word or two successive syllables that are semantically unrelated), we still found a significantly higher error rate in this case than all the cases in the ≤1SYL condition for 6-Hz syllabic rate SUSs (in all pairwise comparison sets, p < 0.05). More direct analysis is displayed as Fig. 4(B): syllable errors averaged across all cases in the >1SYL condition in both 4- and 6-Hz syllabic rate SUSs are significantly higher than those in the ≤1SYL condition (p < 0.0001). These results thus indicate that subjects were able to have relatively good performances in understanding Mandarin spoken sentences with no more than one syllable as a basic unit compared to more than one syllable as a basic unit.

3. Discussion and Conclusion

Through Exp. 1 and Exp. 2, three main results are obtained in the present study: (1) the optimal range of the syllabic rates for efficient Mandarin speech reception (at around 3–8 Hz) is aligned with that of speech normally produced in daily communications (around 4–8 Hz) [2]; (2) the (1/2)SYL condition has generally lower syllable errors than the 1SYL condition (except for the 8-Hz syllabic rate SUSs as shown in Fig. 2(A) (the ×120 cases)); (3) errors of the ≤1SYL condition are significantly lower than those of the >1SYL condition.

According to the first result, we suggest that the optimal range of syllabic rates for syllable intelligibility of the sentences, which is aligned with the normal Mandarin production data, reflect the close relationship between speech perception and production. It is interesting that this seems to be consistent with previous neurophysiological studies [11][12] which suggest that both speech perception and production processes share similar brain areas responsible for mouth and tongue movements. Such areas also possess the 3–8 Hz neural oscillations [5] that are thought to be responsible for the syllable level processing [3]. Furthermore, our present finding that intelligibility is modulated by syllabic rates enhances the idea of viewing syllabic rates as the descriptive index of speech rhythms in Mandarin speech.

The second and third results reveal the syllable and sub-syllable processing in understanding Mandarin continuous speech, namely, Mandarin speakers would preferentially decode spoken sentences into correct syllables with no more than one syllable as the basic perceptual unit. And in this processing, half a syllable would tend to serve as a more fundamental primitive than one syllable.

Our results are also consistent with the hypothesis [3] that there may exist ‘integration time windows’ in the auditory cortex at syllable and sub-syllable levels (detailed introduction in Section 1). Furthermore, the results obtained here may infer that the human brain is able to ascertain how much linguistic information (e.g.: less than one syllable, one syllable or more than one) is loaded within each set of ‘integration windows’ during Mandarin speech perception. In other words, this capability is ultimately associated with a fundamental issue that how human beings recognize syllable boundaries in continuous speech and sequence syllable or sub-syllable segmentations into a serial linguistic order.

It is also possible that the syllable-timing of Mandarin [6][7] may play a role in obtaining the current results. Due to less variation in syllable duration in Mandarin than stressed-timed languages, some particular cases of periodic silence insertions thus perceptually highlight the senses of syllabic rhythms of the sentences (such as the 1SYL condition, where the boundary between each Acoustic Unit is approximately aligned with the syllable boundaries, making a sentence sound like a string of syllables burst out one after another). This may therefore reinforce the importance of syllabic rates for intelligibility of Mandarin. It should be a further question whether similar results or such 3–8 Hz salience can be extended to stress-timed languages, in which durations between stressed and unstressed syllables are of great disparities, like English [13].

In conclusion, by means of the signal synthesis methodology in the previous study [1], our present study found a clear relationship between speech intelligibility and syllabic rates, consolidating the close association between speech perception and production in Mandarin. We then depicted the importance of syllable and sub-syllable as processing units in understanding Mandarin spoken sentences. Moreover, referring to the previous points of view [3][4][5], we proposed that such optimal syllabic rates and processing units in Mandarin speech perception is possibly associated with the corresponding neural activities in the human brain.

4. Acknowledgements

The work was partially supported by grants from the SIAE of CUHK (BME-8115020), from the RGC of Hong Kong (CUHK DG: 2050461), and from National Natural Science Foundation of China (11074267).

5. References