VISUAL INFORMATION AND THE PERCEPTION OF PROSODY

Haruo Kubozono¹, Shosuke Haraguchi²
Kobe University¹, Meikai University²

ABSTRACT
This paper reports the results of a phonetic experiment carried out to explore the potential importance of visual information for the perception of Japanese prosody and its relationship with phonological structure. It specifically examines whether and to what extent native speakers of Japanese are able to accurately perceive the temporal structure of their native language on the sole basis of visual information. It has been found that the word-final contrast between short and long vowels is totally invisible to the native speakers although other types of temporal differences characteristic of monomoraic and bimoraic syllables are readily visible as a crucial distinction in the mora-timed language. Interestingly, essentially the same confusion occurs in Japanese phonology, where word-final long vowels tend to be shortened and neutralized with their short counterparts.

1. INTRODUCTION

It is well known that visual information provides useful cues in the perception of speech (generally known as the McGurk Effect). However, research on bimodal perception has centered largely on segmental distinctions, e.g. [ba] vs. [ga], and little is known about the role of visual information in the perception of prosody. Moreover, it is not clear at all how facts about bimodal perception relate to phonological structure.

With this background, this paper is concerned with the question of whether or not native speakers of Japanese are able to accurately perceive temporal differences solely on the basis of visual information, specifically differences between monomoraic and bimoraic syllables in existing words: e.g. /sa.do/ 'Sado Island' vs. /saa.do/ 'third'; /ka.ko/ 'past' vs. /ka.ko/ 'bracket'; /fa.ni/ 'valley' vs. /fa.nin/ 'other people'; /kai.gai/ 'picture' vs. /kai.gai/ 'foreign land'. Since the mora is the basic temporal unit in the prosodic system of Japanese, these pairs of words involve differences in rhythmic structures as well as phonological length. Confusion of these pairs of words would mean a neutralization of monomoraic and bimoraic syllables which, in turn, would suggest partial destruction of the mora-based prosodic system.

2. EXPERIMENT

2.1. Method

The 'perception' experiment to be described here was designed to explore the McGurk Effect with Japanese listeners, i.e. to address the question of what extent listeners rely on visual cues when they actually perceive speech in daily communication. Rather, they were carried out to see what listeners could (or could not) 'hear' from the visual information alone when they are deprived of auditory stimuli. These experiments, therefore, do not allow us to look directly into the mechanism of visual perception. Yet, the results of these experiments, therefore, do not allow us to look directly into the mechanism or strategies of speech perception. Hence, we will see that 'visual perception' has a close bearing on phonological phenomena, and is indeed constrained by the same principle of temporal neutralization governing phonological structure.

Test words
The experiment featured 30 test words consisting of the 15 pairs of words summarized in Table 1.

<table>
<thead>
<tr>
<th>W-medial contrast</th>
<th>W-final contrast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moraic Nasal</td>
<td></td>
</tr>
<tr>
<td>ama--aNma</td>
<td>aNma--aNmaN</td>
</tr>
<tr>
<td>taniN--taNniN</td>
<td>tani--taniN</td>
</tr>
<tr>
<td>Short/long V</td>
<td></td>
</tr>
<tr>
<td>sado--sado</td>
<td>sado--sadoq</td>
</tr>
<tr>
<td>buru--briju</td>
<td>rubi--rubij</td>
</tr>
<tr>
<td>okasaN--okagaN</td>
<td>kyoozyu--kyoozyug</td>
</tr>
<tr>
<td>obasaN--obagaN</td>
<td></td>
</tr>
<tr>
<td>Moraic obst</td>
<td></td>
</tr>
<tr>
<td>kako--kakko</td>
<td>syutyo--syutyo</td>
</tr>
<tr>
<td>syutyo--syutyo</td>
<td></td>
</tr>
<tr>
<td>Diphthong</td>
<td></td>
</tr>
<tr>
<td>kaga--kaiga</td>
<td>kaiga--kaigaaj</td>
</tr>
</tbody>
</table>

Table 1  List of test words

These pairs consist of two words contrasting in the presence or absence of one of the four types of non-syllabic moras: moraic nasal (hatsuon), moraic obstruent (sokuon), the second half of a long vowel, and the second half of a diphthong. The contrast in the presence or absence of these moraic elements occurs either word-medially or word-finally, except the moraic obstruent, which can occur only word-medially. The accentuation of the pairs in Table 1 were properly controlled for so that the two members of each pair have an identical accent pattern (with the exception of two pairs, /obasaN/ (unaccented) -- /obaasaN/ (accented) and /okasaN/ (initially accented) -- /okaasaN/ (medially accented)).

Stimuli
The 15 pairs of words, or the total of 30 test words, were randomized and numbered. They were read in isolation (i.e. not in a carrier sentence) by a native female speaker of Tokyo
Japanese, and were simultaneously recorded with a digital video camera. On the basis of this recording, the following three kinds of stimuli were produced: (i) visual stimuli (video playback without sound), (ii) auditory stimuli (cassette tape playback), and (iii) audio-visual stimuli (video playback with sound).

Subjects

The following two groups of subjects participated in the perception experiment: 46 native adult speakers of Japanese who have no reported hearing problem, and twelve hearing impaired adult speakers of Japanese. Ten of the second group of subjects are diagnosed as Grade 2 in the scale of hearing impairment, which means that they can hear virtually nothing in natural conversation. All of the twelve hearing impaired subjects use the Japanese sign language in daily communication; in addition, some have had training in lip reading and use this communication mode as well.

Task

Hearing impaired subjects participated only in the visual test, where they were presented only with the visual stimuli. Hearing unimpaired subjects participated in all of the three tests, all taking the visual test first of all and then the auditory or audio-visual test. In each type of test, the subjects were asked to appropriately choose between two similar words after the playback of each test word. The two words were printed on an answer sheet for a total of 30 responses.

2.2. Results and Analysis

As expected, hearing unimpaired subjects showed very good performance in the auditory and audio-visual tests. None of the subjects, in fact, made any error for any word pair in the audio-visual test.

The same group of subjects gave almost perfect scores in the auditory test, too. They did not make any error in distinguishing between long and short vowels, either in word-medial or word-final positions. Some errors were attested with the moraic nasal and the diphthong only when they appeared in word-final positions. However, the subjects’ performance remains very high: the percentage of correct answers was 97% for the word-final moraic nasal and 94% for the word-final diphthong. Overall, hearing impaired subjects had little or no difficulty in distinguishing between bimoraic and monomoraic syllables in both the audio-visual and auditory tests.

However, a substantially different result was obtained from the visual test. Table 2 summarizes the results of this test with the two groups of subjects: (A) hearing unimpaired Japanese and (B) hearing impaired Japanese. Raw scores are given for each type of non-syllabic mora in each of the two positions in the word: ‘165/184’, for instance, means that 165 answers were correct out of the total of 184. Figures in the brackets denote the percentage of correct answers.

<table>
<thead>
<tr>
<th>Subject group</th>
<th>(A)</th>
<th>(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W-medial W-final</td>
<td>W-medial W-final</td>
</tr>
<tr>
<td>Moraic nasal</td>
<td>165/184 168/184</td>
<td>41/48 42/48</td>
</tr>
<tr>
<td></td>
<td>(90%) (91%)</td>
<td>(85%) (88%)</td>
</tr>
<tr>
<td>Short/long Vowel</td>
<td>359/368 141/276</td>
<td>94/96 31/72</td>
</tr>
<tr>
<td></td>
<td>(98%) (51%)</td>
<td>(98%) (43%)</td>
</tr>
<tr>
<td>Moraic obstruent</td>
<td>171/184 ---</td>
<td>46/48</td>
</tr>
<tr>
<td></td>
<td>(93%) (96%)</td>
<td></td>
</tr>
<tr>
<td>Diphthong</td>
<td>84/92 78/92 24/24</td>
<td>22/24</td>
</tr>
<tr>
<td></td>
<td>(91%) (85%) (100%)</td>
<td>(92%)</td>
</tr>
<tr>
<td>&lt;Total&gt;</td>
<td>779/828 387/552 205/216</td>
<td>95/144</td>
</tr>
<tr>
<td></td>
<td>(94%) (70%) (95%)</td>
<td>(66%)</td>
</tr>
</tbody>
</table>

Table 2  Scores of the visual perception test (%)

However, a substantially different result was obtained from the visual test. Table 2 summarizes the results of this test with the two groups of subjects: (A) hearing unimpaired Japanese and (B) hearing impaired Japanese. Raw scores are given for each type of non-syllabic mora in each of the two positions in the word: ‘165/184’, for instance, means that 165 answers were correct out of the total of 184. Figures in the brackets denote the percentage of correct answers.

The results summarized in Table 2 indicate that the two groups of subjects generally achieved fairly good performance in the visual perception test, too. Although not as good as in the auditory or audio-visual tests, the average percentage of correct answers in this test is over 80% for each of the subject groups, which is much higher than the chance level of 50%. This suggests that visual information contains very useful cues for perceiving temporal differences such as those found in the test words.

On the other hand, the results in the same table indicate that visual information was not useful at all in distinguishing between pairs that contrasted in vowel length word-finally: e.g. /sa.do/ vs. /sa.doo/. Both hearing impaired and unimpaired subjects were able to correctly distinguish between these words only at a chance level (p> .2 for each subject group)--hearing impaired subjects were not any better than hearing unimpaired subjects--whereas they were perfectly good at distinguishing long/short vowels in word-medial positions, e.g. /ku.do/ vs. /ku.doo/. This suggests that for native Japanese subjects, vowel length is literally `invisible’ in word-final positions although it is perfectly ‘visible’ in word-medial positions. This type of position-dependent confusion occurs only with words contrasting in vowel length (short vs. long), and not with words contrasting in terms of the presence or absence of a moraic nasal, diphthong, or geminate obstruent, e.g. /ta.ru/-/ta.nin/, /kai.ga/-/kai.gai/, /ka.ko/-/kaak.ko/.
In sum, Japanese subjects cannot ‘see’ the difference between long and short vowels in word-final positions while they can distinguish the phonological contrast in word-medial positions. This is true not only of hearing unimpaired subjects but of hearing impaired subjects as well, who are expected be more skillful in lip-reading than the first group of subjects.

This result may not be very surprising since long and short vowels, as their names suggest, differ primarily in duration or timing. In contrast, presence or absence of moraic nasals and diphthongs involves a segmental change (or a change in formant structure in acoustic terms) as well as a change in duration. Given this orthodox idea about the phonetic nature of long/short vowels, it can be predicted that contrasts in vowel length can be seen in word-medial positions but not as clearly in word-final positions. Namely, the timing between the vowel and the following material gives a useful cue to the duration of the vowel but this cue is not available if the vowel appears in word-final positions. However, the results of the visual perception experiment turn out to be very interesting if we consider the phonological tendency to be discussed in the next section.

3. PHONOLOGICAL PATTERNS

Since the distinction between bimoraic and monomoraic syllables and the resultant preservation of moraic structure are vital for mora-timing as against syllable-timing, we would not expect to see these two types of syllable structures neutralized in any context. However, careful examination of various phonological phenomena in Japanese—both diachronic and synchronic—reveals that the underlying moraic structure is subject to change under a certain context. Specifically, long vowels tend to shorten in word-final positions and consequently become indistinguishable from their short counterparts. Such a position-dependent neutralization occurs in a wide range of phonological phenomena from vowel coalescence to some morphophonological processes.

3.1. Historical Change

When we look at the phonological history of Japanese, we find that long vowels tend to shorten word-finally, particularly in bisyllabic Sino-Japanese compounds (Kubozono, in press). (1a) gives typical examples where vowel shortening has established itself. On the other hand, it is difficult to find examples like the one in (1b) in which long vowels shortened in a word-medial position.

(1) Sporadic shortening (established patterns)
   a. ai.so ai.so ‘sociability’
   nai.syo nai.syo ‘a secret’
   rak.kyo rak.kyo ‘a scallion’
   sen.zyu sen.zyu ‘Senju (placename)’
   b. sa.yo.na.ra sa.yo.na.ru ‘good-bye’

The words in (2) represent an ongoing phonetic change whereby long vowels tend to shorten in word-final positions. Again, long vowels do not shorten in word-medial positions.

(2) Ongoing phonetic change
   hon.tyoo hon.to ‘true’, nyoo.bo ng yoo.bo ‘wife’
   men.doo men.do ‘trouble’, gak.ko gak.ko ‘school’
   sen.se sen.se ‘teacher’, o.ben.to o.ben.to ‘lunch’
   ari.ga.to ari.ga.to ‘thank you’,
   kak.ko kak.ko ‘appearance’

It is worth noting here that this kind of temporal neutralization is not generally observed with bimoraic and monomoraic syllables involving the other types of non-syllabic moras described in Table 1 (see Kubozono, in press, for a couple of exceptions to this).

Vowel shortening also affected long vowels resulting from independent historical changes such as vowel coalescence. Vowel coalescence is a phenomenon that amalgamates vowel sequences into a vowel of a single quality. As shown in (3b), this segmental change did not generally affect moraic structure so that the mora length of the original words remained intact. The only exception to this is the case where this segmental change occurred word-finally. In the examples in (3a, b), for instance, the long vowel resulting from vowel coalescence became short in the word-final position, but not in the word-medial position. Note that this kind of word-final shortening did not take place in monosyllabic words, as shown in (3c). This can be attributed to the general constraint on word minimality which militates against monomoraic outputs.

(3) a. teu.tyoo tyoo tyoo
   tyoo.tyoo tyoo.tyoo, tyoo.tyoo ‘butterfly’
   b. kau.be ko be ‘Kobe’
   c. neu nyoo *nyo ‘urine’

3.2. Synchronic Processes

The historical rule of vowel coalescence remains more or less productive in contemporary Japanese. Typically, this rule creates vowel alternations between a diphthong and a monophthong according to speech style: Diphthongs in careful speech tend to become monophthongs in casual speech (see Kubozono 1999a for the basic rule behind this segmental change). The resulting monophthongs are generally lengthened as if to preserve the original mora length, as illustrated by the adjectives in (4).

(4) Vowel coalescence in Japanese adjectives
   a. ita-i itee ‘painful, Ouch’
   b. sugo-i sugee ‘great’
   c. atu-i atti ‘hot’

In very casual (and fast) speech, however, this lengthening optionally fails to apply—or, equivalently, lengthened vowels are again shortened. Thus we have such forms as /ite/ and /suge/ for /ite/ and /suge/. Interestingly, this shortening occurs only at the end of the prosodic word, and is invariably blocked when the
word forms a larger prosodic word with a following clitic-like element such as the sentence-final particle na. Consequently, such forms as /ite-na/ and /suge-na/ are unacceptable while their counterparts with a long vowel, i.e. /itee-na/ and /sugee-na/, are perfectly acceptable. This symbolizes the contrast between word-final long vowels and their word-medial counterparts with respect to vowel shortening.

The peculiar behaviour of word-final long vowels is also observed in the morphological process of compound clipping (or truncation) in Japanese. As illustrated in (5) below, this process most typically involves taking the initial two moras of each component word to produce a four-mora word, or a word consisting of two bimoraic feet (\(|\)). This process is basically independent of syllable structure, so that any syllable combination is possible as long as four-mora outputs are produced (Ito 1990, Kubozono 1999b).

(5) po.ket.to mon suu.taa [po.ke][mon] `Pokemon, pocket monster'

However, this two-foot symmetry is broken when the second component word begins with a heavy syllable containing a long vowel. In this case, the second component usually leaves just the initial mora, i.e. the long vowel is shortened, to produce a three-mora output:

(6) te.re.hon kaa.do [te.re]ka `telephone card'  
    ba.su.ket.to syuu.ru [bas]syu `basketball shoes'  
    dan su paa.tii [dan]pa `dance party'  
    kan nin gu pee paa [kan]pe `a crib'

In contrast, long vowels are not generally shortened if they come from the first component word, as shown in (7a). They are occasionally shortened as in (7b), but in such a case the following material usually fills in the empty mora slot.

(7) a. waa.do pu.ro-se.saa [waa][pu.ro] `word-processor'  
    kaa na.bi gee.syon [kaa][na.bi] `car navigation'  
    zii.zu pa.ni tu [zii][pan] `jeans'  
    b. paa so na.ru kon pkyuu.taa [pa.so][kon] `personal computer'  
    suu paa kon pkyuu.taa [su.pa][kon] `super computer'

In sum, long vowels display different patterns in truncated compound nouns depending on whether they appear word-medially or word-finally. Vowel shortening occurs typically in word-final positions in this type of words just as in the other types of words we have hitherto seen. This peculiar behaviour of long vowels is also contrastive with the behaviour of other non-syllabic moras: moraic nasals, for instance, are almost never deleted even if they appear word-finally (e.g. /po.ke.mon/ in (5)). This suggests that neutralization of bimoraic and monomoraic syllables is strictly restricted to those contrasting in word length word-finally.

4. CONCLUDING REMARKS

In this paper we have examined the linguistic context where the temporal difference between bimoraic and monomoraic syllables becomes invisible or neutralized both in the visual perception of speech and in Japanese phonology. While the distinction between these two types of syllables is indispensable for the mora-based rhythm of Japanese, they are sometimes neutralized. However, this type of temporal neutralization is highly restricted: it only occurs in word-final positions and, moreover, only affects the contrast between long and short vowels.

In the perception of speech with visual information but without sound, word-final contrasts in vowel length are literally invisible to native Japanese subjects, hearing impaired and unimpaired alike. Essentially the same confusion or neutralization is observed in a wide range of phonological phenomena in Japanese.

It is not clear yet how phonological structure and the visual perception of speech are interrelated with each other, e.g. whether phonology constrains the way we ‘see’ speech, or vice versa. However, it is highly interesting that temporal neutralizations in the two areas are constrained in the same way: they occur only in word-final positions and only with the contrast in vowel length. This reveals a striking correspondence between phonological invisibility and visual invisibility, that is, that what is invisible in phonology is also invisible in visual perception, at least as far as temporal structure is concerned.

5. REFERENCES


Acknowledgements: I am grateful to Keiko Ishii for her assistance in conducting the experiments, and to Tadahisa Kondo, Michin affected Matsui and Akira Ujihira for their assistance in making a statistical analysis of the data. I benefited from the discussions with many people, particularly Reiko Mauza, Akiko Hayashi, Reiko Yamada, Hiroaki Kato, Minoru Tsuzaki and Kiyoshi Honda. The work reported herein was supported by research grants from the Japanese Ministry of Education, Science, Sports and Culture (Nos. 10044010 and 12011102).