
Masako Fujiu and André-Pierre Benguerel

Abstract: A Computer-Assisted Speechreading Training system for Japanese (CAST-J) has been designed to train adults with an acquired hearing loss to improve their communicative ability with an effective use of visual speech information. Adults with cochlear implants are the primary target population for CAST-J, as long as they the auditory channel as their primary source of information for speech perception.

The lessons of CAST-J are organized around the concept of viseme. Fifteen visemes of Japanese have been selected for this study: five vowel, seven consonant, and three haku (mora) visemes. There are sixteen lessons, each of which includes sections designed (1) to review previously taught visemes; (2) to practice the recognition of a new viseme; (3) to practice new and old visemes by the tracking method; and (4) to recap the lesson. Two additional lessons are available for pre- and post-training comparison.

The hardware required by CAST-J includes a microcomputer, a video disc player, a high resolution video monitor, and a printer. The system, with its interactive environment, allows the instructor and/or the trainee to select the appropriate speaking rate (fast or slow) for the video, to practice speechreading with or without audio signal, and to skip the optional lessons or part of a lesson, depending on the level of the trainee.

Key words: speechreading, viseme, computer-assisted, cochlear implant, aural rehabilitation

I. INTRODUCTION

A Computer-Aided Speechreading Training system (CAST) has been developed by Pichora-Fuller and her colleagues as a clinical and research tool to train and assess the speechreading abilities of adults with a mild-to-moderate acquired hearing loss. At present, English and French versions of CAST have been implemented. The purpose of this study is to outline the design and the lessons of a computer-assisted speechreading training program for Japanese (CAST-J), considering more particularly patients with cochlear implants as potential trainees. The actual video-recording of the lessons and training of the subjects are not discussed in the present paper.

II. SPEECHREADING

1. Speechreading and viseme

Speechreading, a term broader than lipreading, refers to the "recognition of a speaker's words by watching his lips, facial expressions, gestures, etc., as well as using closure, rules of language, contextual cues, etc." (p. 1087). Besides the movements of the visible articulators, the receiver integrates various environmental and contextual cues. Speechreading and auditory information are known to work synergistically in speech perception, especially in hearing impaired individuals. Manner of articulation and voicing contrast are known to be better
perceived through the auditory channel, while place of articulation fares better through the visual channel\(^{8}\).

'Viseme' is an important concept in speechreading. The term 'viseme', according to Fisher\(^8\), refers to any individual and contrastive unit of visually perceived speech. It is conceptually analogous to the term 'phoneme', a contrastive unit of auditorily perceived speech. All the instances of the phonemes /p/, /b/ and /m/ have labial articulations which are visually indistinguishable, thus they constitute a viseme, the labial viseme. Each member of a viseme is called an allophone. In this study, back slashes are used to denote visemes and angular brackets are used to denote allophones. For example, \(<p>\), \(<b>\) and \(<m>\) are allophones of the viseme \(\langle p \rangle\). Although the number and the definition of visemes are still controversial, there is a generally accepted hierarchy of visemic groups, from well-defined to less well-defined visemes. In English, the visemes \(\langle p \rangle\), \(\langle t \rangle\), \(\langle k \rangle\), \(\langle s \rangle\), \(\langle r \rangle\), \(\langle l \rangle\), \(\langle f \rangle\), \(\langle m \rangle\), \(\langle n \rangle\), \(\langle u \rangle\) have been shown experimentally to be visually dominant, i.e. they are easily recognized in all phonetic environments, while \(\langle t \rangle\), \(\langle k \rangle\), \(\langle s \rangle\), \(\langle r \rangle\), \(\langle l \rangle\), \(\langle f \rangle\), \(\langle m \rangle\), \(\langle n \rangle\), \(\langle u \rangle\) are less dominant or prone to variation\(^9\). The CAST-J lessons are constructed around the notion of viseme.

2. Speechreading and cochlear implantation

The importance of speechreading in aural rehabilitation has grown with the popularity of cochlear implantation for post-lingually deaf adults. The data reported throughout the world thus far indicate that the improvement of hearing in the patients with cochlear implants appears to be encouraging: some patients have been reported to acquire the ability to use the telephone\(^{11,12}\), hearing with a cochlear implant is described as superior to hearing with a hearing aid by patients who have undergone implantation\(^12\). It has also been reported that some people do well in open-set speech recognition tests, without lipreading\(^{11,13}\). Nevertheless, the majority of patients benefit most from the combination of auditory and visual inputs\(^{11,14-16}\).

The patients who have undergone cochlear implantation in Japan are all adults with an acquired hearing loss. It is suggested that the ability to lipread be one of the criteria for selecting the candidates for implantation. The data for six subjects studied by Fukuda and her associates\(^14\) exhibited enhanced speech perception three months post-operatively, when auditory and visual cues were combined, as compared with auditory cues only or visual cues only. Their data indicates that post-lingually deaf adults with cochlear implants use auditory input as a primary source of speech information, and they can increase their communicative ability a great extent by using supplementary visual cues.

3. Evaluation and training

Currently, several tests for speechreading evaluation are available for English\(^{17-24}\). For Japanese, as far as the development of assessment materials is concerned, very few studies are found in the literature. In the 1970s, Tanaka and his associates developed a video system to assess speechreading ability with the goal of using the assessment results for an adequate hearing aid fitting, primarily for hearing-impaired children\(^{25,26}\). Besides this system, a video system has been developed by Fukuda\(^27\) for assessing the speechreading ability of adults with cochlear implants.

Recently, computer-assisted devices have become increasingly popular in aural rehabilitation. CAST is one of those systems; it has been developed for speechreading evaluation as well as for training. In speechreading training, discourse tracking is one of the most widely used methods. The CAST lessons were constructed in accordance with the principles of the tracking method: face to face interaction, with feedback between sender (in this case the videotape) and receiver (trainee). The investigation of the effectiveness of the CAST program has found a greater improvement of speechreading ability in the experimental group, consisting of eight normal-hearing adults who had been trained with CAST, than in the control group, consisting of eight normal-hearing adults who had not had any speechreading training\(^28\).

III. TARGET POPULATION

The premises of the original version of CAST are that speechreading is "a communication task which entails three skills: visual speech perception, use of
linguistic redundancy, and effective use of feedback between message sender and receiver"\(^1\). The aim of CAST is to improve speechreading, a supplement to auditory perception of speech which can be affected by various environmental factors such as background noise, speaking rate, familiarity with the topic, etc. The potential users of CAST, therefore, are those who use the auditory channel as the primary means of communication, and the visual channel as a supplemental one.

CAST-J will consider cochlear implant patients as primary potential trainees for three reasons: (1) ability to lipread is considered to be an important criterion for potential candidates to the surgery, (2) adults who have undergone a cochlear implantation have been found to have relatively good auditory perception of speech, thus they can use speechreading as a supplementary medium for speech perception, and (3) past research has shown that these patients greatly benefit from speechreading training\(^4,16,20\).

IV. THE SYSTEM

A diagram of the CAST-J system is shown in Figure 1.

The hardware components required by CAST are: (1) a microcomputer including a hard disk, a colour graphics card and a keyboard; (2) a high-quality colour video monitor (14"); (3) a video disc player; (4) an amplifier and a speaker, and (5) a printer. For the English version of CAST, an IBM PC-AT microcomputer with 512-kb RAM, BCD Videolink RS232 video controller, Sony PVM 1271Q TV monitor (14"), Sony SLO 325 video cassette recorder, and Epson FX 286 printer were used. For CAST-J, a hardware setup functionally equivalent to that for the English version of CAST will be used. The main difference is that CAST-J replaces the video cassette player by a video disc player, which provides a faster access time whenever the user wants to jump from one place to another.

The software components of CAST-J system are: (1) an operating system; (2) a set of CAST-J programs written in a high-level programming language (C); (3) a set of CAST-J lessons consisting of the spoken materials recorded on video discs, and of the lesson texts stored on the computer with the programs.

Further information regarding the operation of the system and the detail of the instructions given to the trainee are available in the owner's manual of the English version of CAST\(^11\) and in the study by Aldham, Rochford and Warling\(^32\).

In Japanese, there are several options regarding the input mode and the feedback mode via the keyboard: kana, romaji, and kanji. For this first set

---

![Figure 1](image-url)  
**Figure 1** Functional Diagram of the CAST System
The five vowel visemes and most of the consonant visemes found in conventional Japanese. Lesson 11 through 16 are optional, i.e. the trainee should be allowed to skip or all of these lessons when the lessons become too difficult for him/her, or when the instructor judges that these lessons are inadequate for the level of skill of the trainee.

2. Lesson structure

Basically, the four sections of each lesson are designed, (1) to review previously learned visemes (old visemes), (2) to practice the recognition of a new viseme, in syllables, (3) to practice old and new visemes by the discourse tracking method, and (4) to recap the lesson materials, with a listing of the trainee's performance statistics.

a) Review of previously learned visemes : The reviewing of old visemes is accomplished by presenting the visemes in visual-only mode, in single-haku format (CV) for the consonant visemes, in two-haku format (VV) for the vowel visemes, and in three-haku format (VCCV) for the haku visemes. For this section, only one allophone is selected for each viseme. An example is given in Table 1. There is no review section for the first lesson and for the two general lessons. Viseme tokens are demonstrated and the trainee is simply asked to identify them. He/she may re-play the demonstration of a particular viseme any number of times.

b) Practice of a new viseme : Practising the recognition of a new viseme is accomplished by presenting pairs of visemes in visual-only mode, namely C1a-C2a single-haku sequences for consonant visemes, and V1a-V2a two-haku sequences for vowel visemes (V1i-V2i for the viseme _a_), where C1 or V1 always stands for the target viseme, while C2 or V2 stands for same or another viseme. An example of the preceding is also given in Table 1. The order of the pairs is random. For the training of the haku visemes, two-and three-haku sequences are contrasted, e.g. anpa-apa (or anpa) for \( \text{N} \); appa-apa (or appa) for \( \text{Q} \); and a : i-ai (or a : i) for \( \text{R} \). The trainee is required to discriminate whether the two members of a particular pair of visemes are the same or different. After all pairs have been judged, pairs responded to incorrectly are repeated until they are judged correctly. The trainee may re
Table 1  An Example of a Lesson-Review and Practice of Visemes in Syllable

LESSON 10

Review

Visual Only Presentation:

aa  pa
ia  wa
ua  ra
oa  ta

Training

Visual Only Presentation:

ka  pya
ka  ya
ka  sha
ka  ra
ka  wa
ka  na
ka  ta
ka  kya
ka  ha
ka  ka
ka  gya
ka  za
ka  ba
ka  cha
ka  pa
ka  rya
ka  da
ka  tsa
ka  sa

Audio-visual Presentation:

ka
ga
ha

play any pair any number of times until he/she feels confident that he/she can discriminate the two visemes.

After the discrimination task, all members of the target viseme are presented audio-visually. The trainee may re-play the demonstration any number of times.

c) Practice of old and new visemes in discourse: practice of old and new visemes in accomplished by presenting one or two paragraphs. Each paragraph (text) is characterized by a high frequency of occurrence of the newly learned viseme. Words selected for the texts are words which have a statistically high frequency of occurrence in Japanese. In this section of the lesson, the trainee is expected to integrate visual information of speech with linguistic redundancy and effective sender-receiver feedback in perceiving speech.

Each text is centered around a given topic, so that all sentences in the text are semantically related. All topics are considered to be within the everyday experience and linguistic level of the target population. The topic of each lesson is provided at the beginning of the text.

CAST-J contains both high- and low-predictability sentences, as well as a variety of syntactic structures and phrase lengths. The text for each lesson is divided into phrases of various length (major strings), which are predetermined by the designer (cf. Table 2). A string in this text means a sequence of characters. Each major string is numbered, and used as member of a set of target words/phrases. Each major string is further divided into words or phrases (minor strings), which are also predetermined by the designer. In the text, minor strings are separated by spacing, while they are indicated by pausing in the text read by the speaker. Although an indication of word/phrase boundaries is not mandatory in Japanese writing, the use of ‘minor string’ is imperative for the computer program to provide better feedback: the computer can compare the target and the response more accurately for short strings of characters than for long ones. The text for each lesson consists of approximately 30 to 50 major strings.

In the first 10 lessons, each lesson has two paragraphs, the first consisting of syntactically simple sentences and the second consisting of syntactically complex sentences. The syntactically simple sentences contain few embedded sentences, While the syntactically complex sentences contain a number
Table 2  An Example of a Lesson–Practice of Old and New Visemes in Discourse

Practice B

<table>
<thead>
<tr>
<th>No.</th>
<th>Sentence</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Higoro</td>
<td>GOMINO SHORI (GARBAGE DISPOSAL)</td>
</tr>
<tr>
<td>2</td>
<td>nanigenaku suteteiru</td>
<td>5I5030 #</td>
</tr>
<tr>
<td>3</td>
<td>gominisuite</td>
<td>4A75E4A5U # 4U4E4E13U #</td>
</tr>
<tr>
<td>4</td>
<td>kanshin–o mottakotoga</td>
<td>501I7I4U14E #</td>
</tr>
<tr>
<td>5</td>
<td>koremadeni arimasuka?</td>
<td>5A87I8 # 0 # 1094A50405A #</td>
</tr>
<tr>
<td>6</td>
<td>Dokode</td>
<td>503E14E7I # A31A4U5A #</td>
</tr>
<tr>
<td>7</td>
<td>ikanaru keiwohete</td>
<td>40504E #</td>
</tr>
<tr>
<td>8</td>
<td>shoriga</td>
<td>703I5A #</td>
</tr>
<tr>
<td>9</td>
<td>okonawarete irurodesho:ka.</td>
<td>0504A2A3E4E # 13U404E705A #</td>
</tr>
<tr>
<td>10</td>
<td>Gominowa</td>
<td>501I7I2A #</td>
</tr>
<tr>
<td>11</td>
<td>daidokorokaraderu</td>
<td>4A14050305A3A4E3U #</td>
</tr>
<tr>
<td>12</td>
<td>namagominohokani</td>
<td>4A1A501I40505A7I #</td>
</tr>
<tr>
<td>13</td>
<td>reizo:koo hajimetosuru</td>
<td>3EI40:500 # 5A71I1404U3U #</td>
</tr>
<tr>
<td>14</td>
<td>o:gata, kogatano</td>
<td>05A4A # 505A4A40 #</td>
</tr>
<tr>
<td>15</td>
<td>denbaisekin,</td>
<td>4E85A4E18 #</td>
</tr>
<tr>
<td>16</td>
<td>kagunadoga fukumaremasu,</td>
<td>5A5U4A405A # 6U5U1A3E1A4U #</td>
</tr>
<tr>
<td>17</td>
<td>Kakuchide</td>
<td>5A5U7I4E #</td>
</tr>
<tr>
<td>18</td>
<td>kaishu:sar eru gomino</td>
<td>5A17U:4A3E3U # 501I40 #</td>
</tr>
<tr>
<td>19</td>
<td>hobo gojuttpa: sentoga</td>
<td>5010 # 507U91A4E8405A #</td>
</tr>
<tr>
<td>20</td>
<td>kanengomide,</td>
<td>5A4E8501I4E #</td>
</tr>
<tr>
<td>21</td>
<td>korerawa</td>
<td>503E3A2A #</td>
</tr>
<tr>
<td>22</td>
<td>kindai dasaret seiso: kojo: ni</td>
<td>5I84A15A4A3E4A # 4EI40: 507:7I #</td>
</tr>
<tr>
<td>23</td>
<td>hakobikomaremasu</td>
<td>5A501I501A3E1A4U #</td>
</tr>
<tr>
<td>24</td>
<td>Gominonakana toetsubunawa</td>
<td>501I404A5A40 # 4E4U182A #</td>
</tr>
<tr>
<td>25</td>
<td>jishakuo tsukatte</td>
<td>7I7A5UO # 4U5A94E #</td>
</tr>
<tr>
<td>26</td>
<td>kaishu:sare,</td>
<td>5A17U:4A3E #</td>
</tr>
<tr>
<td>27</td>
<td>shigen no sairyo: kaga</td>
<td>7I5E840 # 4A1I7I05A4A3E #</td>
</tr>
<tr>
<td>28</td>
<td>hakarareteimasu</td>
<td>5A5A3A3E4EI1A4U #</td>
</tr>
<tr>
<td>29</td>
<td>Chikagorouya</td>
<td>7I5A50302A #</td>
</tr>
<tr>
<td>30</td>
<td>shinkai hatsusareta</td>
<td>7I85A15A4U4A3E4A #</td>
</tr>
<tr>
<td>31</td>
<td>sozai ga ichihayaku</td>
<td>404A15A # I7I5A7A5U #</td>
</tr>
<tr>
<td>32</td>
<td>kateiseikutsu: hairikomi,</td>
<td>5A4EI5A4U7I # 5A3I5I01I #</td>
</tr>
<tr>
<td>33</td>
<td>atarashi: ko: gaito</td>
<td>4A3A5I7I5:50:5A40 #</td>
</tr>
<tr>
<td>34</td>
<td>narunode wanaikato</td>
<td>4A3U40E2A4A15A40 #</td>
</tr>
<tr>
<td>35</td>
<td>kenensareteimasu</td>
<td>5E4A84A3E4EI1A4U #</td>
</tr>
<tr>
<td>36</td>
<td>Konoyo: ni</td>
<td>504070:7I #</td>
</tr>
<tr>
<td>37</td>
<td>gominikanshilemo,</td>
<td>501I7I5A87I4E10 #</td>
</tr>
<tr>
<td>38</td>
<td>kaikezushinakerebanaranai</td>
<td>5AI5E4U7U4A5E 3E1A4A3 A4AI #</td>
</tr>
<tr>
<td>39</td>
<td>kadaiga</td>
<td>5A4A15A #</td>
</tr>
<tr>
<td>40</td>
<td>kanari arunodesu</td>
<td>5A4A3IA3U404E4U4 #</td>
</tr>
</tbody>
</table>

of embedded sentences. In the next six lessons, each lesson has one single paragraph with sentences of various syntactic structures. Lesson 17 and Lesson 18 are also one-paragraph lessons. The paragraphs used in these two lessons are considered to be visemically well balanced, and to be equivalent in terms of occurrence of each viseme, syntactic complexity, and word familiarity. They are suitable for pre- and post-training comparison.

In CAST–J in the third section of the lesson (practice of old and new visemes in discourse), the speaking rate of the texts (both at paragraph level and phrase level) can be selected. Two readings of each text are recorded, at different speaking rates: slow
normal speech and fast normal speech. The trainee or the instructor selects the appropriate rate for each lesson. The modality of presentation can also be selected. Visual-only and audio-visual presentation are available. The trainee or the instructor can select the modality adequate for each lesson. The presentation of lessons can be made with or without amplification, and/or with or without background noise. All responses and their corresponding feedback are always displayed on screen (on a worksheet), throughout the lessons. Compared with the live face-to-face procedure, this display is helpful in two ways: (1) it eliminates the possible ambiguity of the response; (2) it eliminates the misunderstanding of feedback by the trainee, who, in the face-to-face situation, may have the same difficulty understanding the spoken feedback as understanding the target phrases. For this particular section of the lesson, there is a ceiling to the number of times that the trainee is allowed to repeat the same target when his/her response is not correct. However, the trainee is given the freedom to skip a phrase and to come back to it later.

3. Romaji-viseme conversion and target response matching

In the tracking procedure used in CAST-J, feedback is provided to the trainee regarding the accuracy of his/her responses. Target string and response string are converted into visemic code, following the rules of romaji-viseme conversion. Vowels are coded with alphabetic characters: A, I, U, E, and O: consonants are coded by numbers: 1 through 7; and the hake visemes \( \overline{N} \), \( \overline{Q} \), and \( \overline{R} \) are coded by 8, 9, and a colon (:), respectively. Feedback is provided on the basis of the visemic match between target and response. The matching algorithm compares response and target, character by character. Visemes which have not yet been
presented, in previous lessons or in the current one, are coded as C if they are consonants, /N/ /Q/, and as V if they are vowels, or /R/. When the trainee fails to identify the existence of a certain viseme, i.e. when he/she does not input any response for the target viseme, the feedback for this viseme is an underscore, regardless of the level of the viseme. A typical interaction in Lesson 8 would appear on the screen as shown in Table 3:

However, when the trainee identifies the presence of a viseme but misidentifies it, i.e. when he/she puts the wrong viseme for the target, the feedback provides the answer for visemes which are not yet learned (C and V), but it only provides an underscore for previously learned and currently learned visemes. Table 4 illustrates this point for a trainee studying Lesson 5:

Thus, feedback is more demanding in later lessons, where the trainee is expected to know a certain number of visemes, than in earlier lessons, where the trainee is expected to identify only a few visemes.

In addition to a simple position-by-position match of the visemic code, a rightward search takes place. If the visemic code for the response does not match the visemic code for the target in the corresponding position, the matching program continues to check the next characters, until the end of the minor string (i.e. until the next word/phrase boundary). If a match is found, (a number or C for a consonant, a letter or V for a vowel), the search re-start from the visemic code which is immediately right of the one matched last. Table 5 illustrates this point for a trainee studying Lesson 9:

Although the matching provided by the system is not always ideal, it can provide feedback which is not too far from that could be given by an actual instructor.

IV. DISCUSSIONS

This study is the first attempt to design a computer-aided speechreading training system for Japanese (CAST-J). In order to implement it and to further refine it, several factors must continue to be investigated.

The first factor is the definition of the viseme, more particularly in relation to coarticulatory effects. In view of the significance of haku, i.e. the strict (C) j V structure in Japanese, it may be desirable to group visemes in accordance with their visemic context. For example, instead of the alveo-dental viseme /t/ assumed in this study, one may favor two separate visemes /t₁/, in those contexts where it is followed by /a/, /e/, or /o/, and /t₂/, in those contexts where it is followed by /i/ or /u/.

The second factor is the influence of sociolinguistic changes on the phonetic inventory of Japanese. For instance, the fricative consonant[f], which is not included in the phonetic inventory of conventional Japanese but is a well-defined visemes in English, has been shown to occur in certain sociolinguistic groups.

The third factor is the adequacy of the input and feedback mode, i.e. the choice of kana or romaji. The advantages and drawbacks of using romaji or kana needs to be considered further in the development of CAST-J.

The fourth factor is the relation between syntactic complexity of the text and speechreading performance. Past research has shown that linguistic complexity affects the speechreading performance. Further research in this area will be helpful for the development of adequate lesson materials and the selection of an appropriate group of population.

The fifth factor is the design of other versions of CAST-J for different populations, for example children with a congenital hearing loss. With a better understanding of speech perception and language development in the above populations, CAST-J will make a contribution to enhance the communicative ability of the hearing-impaired individuals.

VII. CONCLUSIONS

A Computer-Assisted Speechreading Training System for Japanese (CAST-J) has been described. The system has been designed primarily for patients with cochlear implants. The organization of the sixteen lessons of CAST-J, which are presented in an interactive manner between the computer-controlled training program and the trainee, is based on the visemic hierarchy and the linguistic
features of Japanese.

Acknowledgement

The authors would like to thank those who have contributed to this work: Kathy Pichora-Fuller, Noelle Lamb, John Nicol, Yumiko Fukuda, and Ikuko Seki. We also would like to express our sincere appreciation to Dr. Hajime Hirose and the staff at the Research Institute of Logopedics and Phoniatrics, University of Tokyo, for offering us an opportunity to complete this submission.

References


21) Sims, D. G. and Jacobs, M.: Speechreading evaluation and the National Technical Institute


Request for reprints:
Masako Fujiu
Department of Communication Sciences and Disorders
Frances Searle Building Northwestern University
2299 Sheridan Road Evanston, Illinois 60208 U. S. A.