INPUT MODALITY EFFECTS IN FOREIGN ACCENT

Duncan J Markham & Yasuko Nagano-Madsen°*

*Dept of Linguistics & Phonetics, Lund University, Helgonabacken 12, 223 62 Lund, Sweden
°Dept of Oriental Studies, Gothenburg University, 412 98 Gothenburg, Sweden

ABSTRACT

A group of higher than average ability learners of the phonetic characteristics of a foreign language attempted to imitate stimuli from Japanese in tasks involving long and complex phrases and shorter minimal pair contrasts. Three information modalities were used (aural, visual, orthographic), introduced progressively, in order to test the subjects' use of different input sources. Their performance is described and the conclusion is drawn that these speakers function largely within an aural modality, but do assimilate external, primarily visual, information, as reflected in rapid improvement in the rhythmic and timing characteristics of the imitations.

1. INTRODUCTION

Studies of foreign accent using imitation of foreign language sounds have generally yielded positive results with regard to the ability of imitators (learners and naive imitators) to produce authentic versions of the foreign stimuli ([1-3]). This paper addresses naive imitative performance, and the differences observed in imitations as a function of the phonetic input modality. Empirical evidence for the usefulness of certain types of training information is found in [4], where a group of subjects given solely auditory information were observed to be less good at producing 'exotic' sounds than a group given explicit articulatory information. Similarly, [5] found that naive learners of Mandarin tones performed better in production when trained first in the perception of the tones, rather than when trained in production and then in perception.

Perception research has yielded at times contradictory results regarding the utilisation of visual information. Although [6] found that Japanese speakers were largely unaffected by visual information when visual and auditory input did not correspond, [7] presents data supporting the integration of multiple input sources in the perception of speech.

Given these observed effects of different training and input modalities, we decided to test the use of information from three input sources in the production of a selection of phrases and words in Japanese. These three modalities were aural, visual, and orthographic. The speakers selected for the experiment were five native speakers of Swedish with demonstrated success in learning the phonetics of a foreign language. By using subjects with a 'talent' for phonetic acquisition, it was hoped that the wide variation in performance in most tests of L2 pronunciation using randomly selected subjects (cf [5]) could be reduced .

The aural material was collected within another research project ([8, 9]), and as such was not explicitly designed for this experiment, but was deemed to be adequate for our present intentions. This material consisted of the words and phrases shown in Table 1. The stimulus material was drawn from a text read at two different speeds by a female native speaker of Osaka Japanese (YN-M) (with some Standard Japanese influences), and from a list of words read by the same speaker. In the PhraseWord task, three phrases from the text were chosen, along with words which appeared in these phrases. In the Contrast task, words containing sounds usually difficult for Swedish learners of Japanese were selected and presented both singly, and as opposed minimal pairs.

The imitators were fitted with lightweight headset microphones and high-quality headphones. After having been familiarised with the stimulus speaker's voice and having heard the two (slow, normal) readings of the text in Japanese, the subjects heard the stimuli for imitation. The learners were not provided with any phonetic representation of Japanese (ie, neither romaji, hiragana/katakana, nor phonetic transcription), thus having to rely

PhraseWord Task			yoko ni	/ jokoni /	
Sutefan wa pairotto de			yokoo ni	/ jokooni /	yoko ni-yokoo ni
Sutefan wa	pairotto de		yokyoo ni	/ jokʲooni /	yokoo ni-yokyoo ni
ofu /ohɯ/ to itta guaidesu			anpan o	/ aNpaNo /	
ofu to	itta guaidesu		onna o	/ oNnao /	
futariwa /hutaɾiwa/ Nihon Kookuu de hataraite imasu			hoteru de	/ hoteɾɯde /	
futariwa	Nihon Kookuu de	hataraite imasu	hanashite	/ hanasite /	
futariwa Nihon Kookuu de			nihon no	/ nihoNno /	
Nihon Kookuu de hataraite imasu			hána ga	hana ga	hána ga-hana ga
			kúmo ga	/ kumoga /	
Contrast Task			kumó ga		kúmo ga-kumó ga
byooin de	/ bʲooiNde /		háshi ga	/ hasiga /	
biyooin de	/ bijooiNde/	byooin de-biyooin de	hashi gá		háshi ga-hashi gá
nyuugaku ga	/ nʲɯɯgakɯga /		kyooryuu	/ kʲooɾʲɯɯ /	
ikka ni	/ ikkani /		raku	/ ɾakɯ /	
ikki ni	/ ikkini /	ikka ni-ikki ni	hatsuga	/ hatsuga /	
ryokoo ga	/ ɾʲokooga /		hazuga	/ hazɯga /	hatsuga-hazuga
ryukku ga	/ ɾʲɯkkɯga /				

Table 1: Stimuli used in the two imitation tasks.

solely on auditory input. In the PhraseWord task, each stimulus phrase was hear three times at each speed, and then the stimuli for imitation were presented. The stimuli were at first single words from the phrase and gradually became more complex (subphrases and then whole phrases). Each stimulus presentation consisted of two instances of the item, which was then imitated by the subject. A total of three presentations were made for each stimulus, yielding three imitations per subject per stimulus. In the Contrast task, subjects only had one opportunity to imitate the stimulus (one presentation).

2. VISUAL AND ORTHOGRAPHIC TESTS

Six months after the aural trial, the speakers were shown a video of the Japanese speaker reading the text and then the individual words and phrases. It was hypothesised that the visual input would provide extra information for the speakers as to the movements necessary to produce some of the sounds. There is some support, though almost entirely theoretical, for the notion of language specific 'settings' which are articulatory configurations typical to a given language [10-12]. If these settings exist, it was hypothesised that the visual information should provide peripheral cues to the nature of the setting for Japanese.

The video was prepared by getting the Japanese speaker to imitate herself, using the original test stimuli used on the five subjects. The video image covered from just below the speaker's chin to just below the speaker's cheek-bones. These imitations — which were very close to the originals — thus recreated the articulatory and tempo information of the original stimuli. The original stimulus recording was then dubbed onto the video tape, synchronised to the imitations by the Japanese speaker. This resulted in a video recording containing the original stimulus presentation and synchronised lip movements. In at least 80% of cases, the timing of the visual information was so close that naive watchers could not readily see any discrepancy in phasing between the visual and audio signals.

A large (30 inch) television set was placed close to the observation window of the recording booth so that subjects had a good and unobstructed view of the entire screen through this window. The audio signal from the VCR was fed into the studio via a headphones connection. In this way, the subjects heard exactly the same stimuli as in the original aural task, but with new, visual, information added.

After completing the two imitation tasks, the recording session was interrupted for approximately 10 minutes, and the subjects chatted with one of the investigators. During this time the investigator asked the subjects how they had found the imitation task with the new (visual) information, and whether they had noticed anything about Japanese which they hadn't thought about in the solely auditory imitation task conducted six months prior. The speakers then watched the instruction video again, but this time with a romaji (Roman character) transcription of the material presented on the television screen along with the visual articulatory information, and completed the imitation tasks once more.

2.1 Speakers' Subjective Reactions

All speakers commented on the fact that Japanese seemed to be articulated with very little lip movement, and with spread lips. Three of the five subjects also thought that the large movements which did occur were vertical *jaw* movements. All said that they had been helped on occasion by the visual information in deciding how a perceptually difficult sound might be articulated. However, not all liked being presented with visual information. One subject thought that the visual signal was more of a distraction than help, yet at the same time admitted that it had been useful once or twice.

The orthographic information was not commented on explicitly by most speakers. One speaker said, however, that she had tried to ignore it as far as possible, as she felt that orthography in general was usually misleading when trying to learn unfamiliar sounds. Another speaker (who had found the visual info a distraction) also felt that orthography was too often a trap, especially as she noticed that the phonetic value of some of the romaji graphemes (or perhaps her perception of the acoustic signal) did not relate to her expectations of the values of Roman characters (this related especially to the /ɾ/ and /ɾʲ/ phonemes, written as <r, ry>).

3. RESULTS AND DISCUSSION

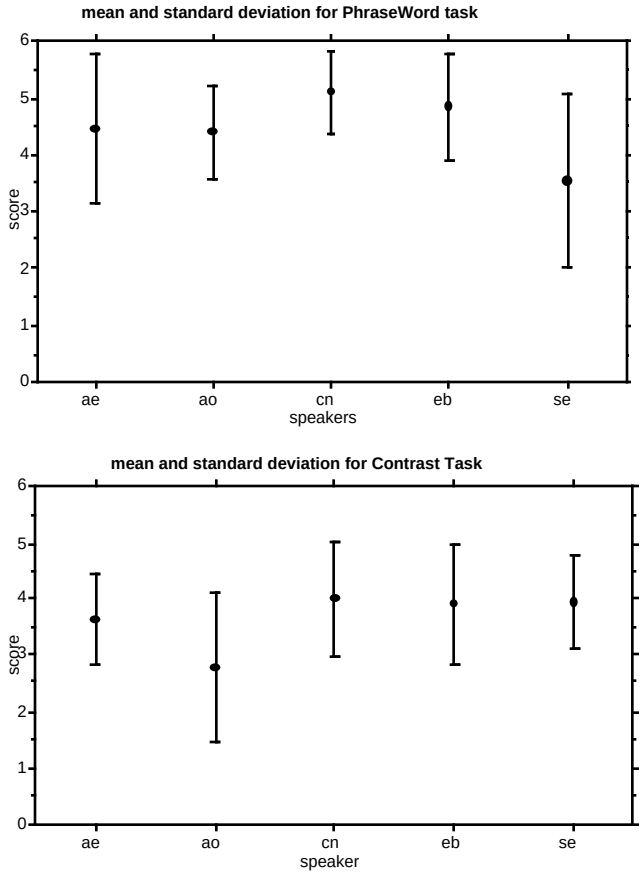
It was our hypothesis that the two input modalities would have different effects on the imitators. It was thought that the visual signal would provide information regarding labiality, and general labial and mandibular positioning and movement. The orthographic information was expected to help speakers in repairing potential segment miscategorisations, especially with regard to the realisations of /ɾ, h [ɸ], and ɯ [ɯ̟]/ and palatalised phonemes. It was also thought that the orthographic representation of phonetic geminates would lead to an improvement in their production. Negative effects were also predicted, with /h/ expected to become labiodental, as romaji uses <f> for the conditioned allophone of this phoneme, /ɾ/ to become more rhotic (approximant or trill-like) than the native Japanese tap realisation, and devoiced vowels to become fully voiced (voiced and devoiced vowels are not differentiated in romaji).

3.1 Imitation Problems and Accent Scores

The imitations in the first test were scored by the speaker of Japanese for degree of foreign accent. A 7-point scale, ranging from "0" (definitely native) to "6" (very strong foreign accent), was used. The imitations were *not* scored for closeness of imitation to the original. As can be seen in Figures 1 (PhraseWord task) and 2 (Contrast task), the obtained scores were not very good, with mean scores ranging between 3.5 and 5.1. For the PhraseWord task, two-tailed paired t-tests of inter-speaker accent scores yielded values of t at $p \leq 0.05$ or better for speaker SE against all other speakers, and for speakers AE and AO against speakers CN and EB. However, in the Contrast task, different results obtained, with AO being the only speaker yielding a t value with $p < 0.05$, against all other speakers. Speakers SE and AO were the only imitators to receive scores of 0 or 1 for any of their imitations in the tasks where their scores differed significantly from the remaining speakers.

The stimuli in the PhraseTask contained the following words and segment sequences found to be produced especially poorly by the subjects:

/**hɯt**aɾiwa, o**hɯt**o/ → [ɸ$^{\Psi}$t] where ɸ was often reproduced as a velar or labiovelar fricative (Swedish ɧ), presumably because the low frequency noise spectra are similar for all three sounds. In addition, the /hɯt/ sequence, where the /ɯ/ is devoiced, sounded tense, as if the two consonants belonged to a single syllable onset, with a short fricative and long stop [ɸt:], whereas the sequence in Japanese is a perceptibly less tense[ɸ$^{\Psi}$t]. This did not occur in



Figures 1 & 2: Mean accent scores for five subjects in PhraseWord task (n=20) and Contrast task (n=35)

ofuto, as the first syllable appeared to be perceived as more prominent, and the fricative was then assigned to the coda of the first syllable.

The /**w**/ in Japanese is usually transcribed as [w], but is in fact an unrounded bilabial approximant similar to [ɤ]. The subjects' productions sounded like [w] in all cases. The triphthong in /g**ɯai**desu/ → [ɤ̆ă̆i] was produced as [waˌi] by all speakers.

In the Contrast task, single words and minimal pairs contrasting features such as palatalisation, quantity, pitch accent, and voicing were imitated. All speakers produced acceptable unpalatalised velar stops, despite the fact that this is usually a problem for Swedish learners of Japanese (Swedish often has prevelar offsets for these stops, whereas Japanese has palatalised and normal velar stop phonemes), whilst all were poor in differentiating between the sequences /bʲooiNde/, /bijooiNde/. The pitch accent contrasts were also very poorly produced for many words. Despite the fact that the Contrast task only allowed one imitation attempt, and emphasised difficult contrasts, the mean scores of all speakers except SE were in the order of approximately one score field better than in the PhraseWord task.

3.2 Input Modality Effects

The first observation must be that there was considerable overall improvement for all speakers from the audio to the visual stage of the experiment. Part of this improvement may be due to a training effect, considering its magnitude, although we were rather surprised. Six months had passed between the two stages, the subjects had not had any further exposure to Japanese, and the subjects had not been aware that there would be any further experiments after the audio stage, so had no reason to want to retain any of the information or representations they had gained in the first stage. However, the *nature* of these speakers was such that they all harboured a curiosity for phonetics and languages, and thus may exceed our expectations for information retention. The most striking improvement was in the quality of the intonational prosody, which obviously cannot be an effect of the visual signal, although it might not be inconceivable that *rhythmic* information reinforced by the visual signal and reflected in the imitations improved the perceived quality of the intonation.

The results of the aural test revealed difficulties in imitating certain consonants. After the addition of visual input, consonant quality improved dramatically for some problem sounds such as /h and w/. The tap realisation of /ɾ/, and the plain/palatalised consonant contrast were problematic for most imitators, and despite predictions of improvement after addition of orthographic information, any change was found to be largely speaker-dependent.

Unexpectedly, clear improvements in vowel quality were found in the visual test, with further improvements, of lesser magnitude, for the orthographic test. These changes largely involved utterance-final vowels in stimuli like *hoterud**e*** and *hana g**a***, and may also have related to vowel quantity or termination. Improvement in vowel quality was negligible for the PhraseWord task.

Improvements probably attributable to visual information were observed for /w/ and /h[ɸ]/ (sounded as if there was decrease in rounding and dentality respectively, and shortening of the latter in /hɯt/), and in the /ɯai/ triphthong. In the latter case, it is proposed that the observable small jaw movement associated with the triphthong was sufficiently different from the excursion posited for the imitators' [waˌi] renditions, that a revision of the production model resulted. Other improvements which may have resulted due to timing/rhythmic information in the signal were generally better vowel quantity and quality (/e/ and /a/ had been too central), though this conclusion can only be speculative.

Several other improvements not obviously attributable to the new information source involved VOT in word initial stops, the production of /ɾ/ as a tap instead of a trill, and the velar nasal realisation of the mora nasal in /aNpa***N***o/.

In the orthographic test, very few systematic improvements were seen. Individual subjects improved for some segments, but there was no obvious regularity across the board. In general it can be said that there was perhaps little room for group improvement after the visual test. The remaining problems were of a largely individual nature. The two most noticeable remaining effects were not attributable to the new orthograhic information, being pitch and vowel quality.

Individual improvement which might be derived from orthographic information involved /h/ for CN, /hoteɾu***d***e/ for SE ([d] instead of [ɾ]), and /ɾ/ for EB. Speakers who had introduced additional syllables in long stimuli due to short-term memory failures (eg, *futari**di**wa*) did not do so in this final test.

Deterioration as a function of the orthographic information occurred in one production of /h/ ([f] instead of [ɸ]), and in two productions of full vowels for devoiced vowels (undifferentiated in romaji). However, some predicted negative effects were not observed, in particular the voicing of devoiced vowels word-internally, the trilling or approximantisation of /ɾ/, and pronunciation of voiceless stops with long lag VOT.

No improvement was observed for final /ɯ/ in *imasu* (a full vowel in Osaka Japanese), which was consistently pronounced as [i] by two subjects. This was particularly unexpected, as /ɯ/ had presented no problems for these speakers in other positions.

3.3 Other Observations

Whilst global pitch was good from the beginning, major improvement in pitch accent was observed for the visual test and to some degree in the orthographic test, but only for the Contrast task, which contained shorter stimuli. Most improvement occurred in words with simple segmental structure. Words containing difficult segmental material barely improved in pitch accent realisation. We draw the conclusion that improvement in pitch is not possible until the segmental characteristics of a stimulus have been processed and, in the view of the imitator, mastered. Modelling efforts are then directed towards pitch, and possibly other prosodic features.

These speakers all showed instances of ignoring input from romaji, where we had expected both positive and negative effects. This usually happened where the subjects regarded the difference between auditory input and expectations from the Roman characters as too great, examples being /r/ which remained tapped or even produced as a stop, and the mora nasal /N/ produced correctly as both [m and ŋ], despite the orthographic <n>. Clearly, the subjects obey their auditory intuitions, this perhaps being a factor in their ability to convincingly imitate other language material to which they have had more exposure.

4. DISCUSSION & SUMMARY

In a study of five Swedish naive imitators' use of different information modalities in reproducing Japanese stimuli, it was observed that the subjects' performance improved generally when visual information was available, but much less so when orthographic (phonological) information was added, despite clear potential for improvement.

We conclude that a visual signal provides information about both specific segments and general timing characteristics of the stimuli which is utilised by the subjects. It is however clearly still the case that they rely primarily on auditory input, and will ignore, sometimes consciously, potentially useful input from other modalities if it does not correspond to their initial modelling of the auditory input.

The most significant effect of the addition of a visual signal in this research was, in our opinion, the improvement in articulatory timing features which are responsible for the rhythmic characteristics of Japanese. Since length is distinctive in consonants and vowels in Japanese, rhythmic relationships are very important, and this area is known to be a problem for foreign learners of Japanese [13]. The imitators had little difficulty in producing acceptable length contrasts, except for vowels in complex stimuli. These (PhraseWord) stimuli were an average of 4 morae longer than the Contrast task stimuli. Swedish also has length contrasts, which presumably helped, and the Contrast stimuli explicitly emphasised many of the target contrasts.

The study gives initial data on issues in input processing and integration. It would be desirable to reproduce this research under more controlled conditions: specifically with three much larger subject groups representing a normal learner distribution, and with each group learning under each of the three different input combinations, thereby eliminating possible training effects.

REFERENCES

1. Neufeld, G., "On the acquisition of prosodic and articulatory features in adult language learning," *Canadian Modern Language Review*, 34, 1978.
2. Locke, J. L., "Short-term auditory memory, oral perception, and experimental sound learning," *Journal of Speech and Hearing Research*, 12: 185-192, 1969.
3. Tahta, S., M. Wood and K. Loewenthal, "Age changes in the ability to replicate foreign pronunciation and intonation," *Language & Speech*, 24: 363-372, 1981.
4. Catford, J. C. and D. B. Pisoni, "Auditory vs. articulatory training in exotic sounds," *Modern Language Journal*, 54: 477-481, 1970.
5. Leather, J., "Perceptual and productive learning of Chinese lexical tone by Dutch and English speakers." In J. Leather and A. James (Eds.), *New Sounds 90, Proceedings of the 1990 Amsterdam Symposium on the Acquisition of Second-language Speech*, pp. 72-97, University of Amsterdam, Amsterdam, 1990.
6. Sekiyama, K. and Y. Tohkura, "McGurk effect in non-English listeners: few visual effects for Japanese listeners hearing Japanese syllables of high auditory intelligibility," *Journal of the Acoustical Society of America*, 90: 1797-1805, 1991.
7. Massaro, D. W. et al., "Bimodal speech perception: an examination across languages," *Journal of Phonetics*, 21: 445-478, 1993.
8. Markham, D., "Investigating imitative ability." In K. Elenius and P. Branderud (Eds.), *The XIIth International Congress of Phonetic Sciences*, Vol. 1, pp. 314-317, Stockholm University/KTH, Stockholm, 1995.
9. Markham, D. J., *Phonetic Imitation and Adaptive Performance*, Lund University Press, Lund, forthc.
10. Honikman, B., "Articulatory settings." In D. Abercrombie et al. (Eds.), *In honour of Daniel Jones*, pp. 73-84, Longmans, London, 1964.
11. Kelz, H. P., "Articulatory basis and second language teaching," *Phonetica*, 24: 193-211, 1971.
12. Wenk, B. J., "Articulatory setting and de-fossilization," *Interlanguage Studies Bulletin—Utrecht*, 4: 202-220, 1979.
13. Imada, S., "Hatsuon no goyoo bunseki no kokoromi." In M. Sugito (Ed.), *Nohongo to Nihongo Kyooiku*, pp 47-71, Meiji Shoin, Tokyo, 1989.