Perceptual Training and the Production of English Consonants by Japanese Learners

Midori IBA

Abstract

This study investigates the impact of intensive perceptual training on the pronunciation of English consonants by Japanese students. I have already examined the effectiveness of using visual cues in the perception of labial/labiodental contrast in a previous study (Iba 2004). According to that study, audiovisual training seemed to be more effective in improving the perception of the labial/labiodental contrast than auditory training. In the study, Japanese learners of English were tested on their perception of the /l/-/r/ contrast in audio, visual and audiovisual modalities, and then undertook ten sessions of perceptual training with either auditory stimuli, natural audiovisual stimuli or audio-visual stimuli with a synthetic face synchronized to natural speech. The /l/-/r/ perception improved in all groups, but learners trained audiovisually did not improve more than those trained auditorily. Auditory perception improved most for 'A training' (audio training) learners, and sensitivity to visual cues improved most for 'AV training'(audio-visual training) learners. The learners' pronunciation of /l/-/r/ improved significantly following perceptual training, with a greater improvement seen for those trained audiovisually with natural stimuli. The present study shows that sensitivity to visual cues for non-native segmental contrasts can be enhanced via perceptual training, and that audiovisual training will be more effective for those phonemic contrasts for which visual cues are sufficiently salient.

1. Perception of L2 sounds

Learning to perceive and produce sounds of a second language can be effortful if these do not occur in the native language or have a different phonological status in the two languages (e.g. Flege 1995; Guion et al. 2000). In addition to segmental errors in perception (e.g. Takata and Nabelek 1990), there can be an increased cognitive load when listening in a non-native language (Tyler 2001). It has also been shown that even highly fluent second language learners who have similar intelligibility to native listeners for speech in quiet sorroundings, show disproportionate reductions in intelligibility when speech is degraded by noise (Mayo et al. 1997).

A number of factors affect the degree of difficulty in acquiring novel phonemic contrasts. One important factor which has received a lot of attention is the relation between the phoneme inventories of the first (L1) and second (L2) language, as our knowledge of the phoneme inventory of the L1 interferes with the acquisition of that of the L2. The Speech Learning Model (Flege 1995) suggests that second-language learners tend to assimilate L2 sound categories to their native sound categories if the categories are phonetically similar. Categories that are phonetically-dissimilar enough from native sound categories are seen as easier to acquire as there is no process of L1 interference, whilst categories which are identical in the two languages are typically perceived and produced without too much difficulty. Best's Perceptual Assimilation Model (PAM), based on a different theoretical framework, known as the Direct Realist Model of Speech Perception (e.g., Fowler 1986), makes detailed predictions of L2/L1 assimilation on the basis of similarities in articulatory gestures between the L2 and L1 sounds. An L2 sound is either 'categorized' as an exemplar of a native phoneme category, 'uncategorized' if it is similar to two or more native categories, or 'nonassimilable' if it is not at all similar to any native category. Both of these theories predict that certain phonemic contrasts will, at the very least, be difficult to acquire within a specific L1/L2 combination. For example, the /l/-/r/ contrast in English has been extensively used in studies of L2 speech perception as it is a difficult contrast to acquire for Japanese learners of English due to the different phonological status of this contrast in Japanese. In Japanese, lateral approximants do not occur, but there is an alveolar flap [β] within the phonemic inventory. The English phonemes /l/ and /r/ tend to both be assimilated to the native alveolar flap, leading to problems in the discrimination and identification of these English phonemes. The /l/-/r/ contrast is particularly difficult for Japanese learners because it is primarily marked by differences in third formant transitions to which Japanese listeners are not particularly sensitive, presumably because F3 transitions do not act as primary acoustic cues for Japanese sound distinctions (Yamada 1995; Iverson et al. 2003).

Another important factor that affects the acquisition of acquiring novel phoneme categories in an L2 is the age of second language acquisition (for a discussion, see Flege 1999). Although there is now little support for the view that there is a critical period of acquisition (Lenneberg 1967) after which the acquisition of new phoneme categories would be impossible, there does appear to be a sensitive period for L2 acquisition, at the phonetic level at least. Indeed, in studies of immigrant L2 learners from a homogenous L1 population, the degree of foreign accent in production (as assessed by native listener ratings) increased fairly consistently with age (e.g. Flege 1998).

Another argument in favour of ongoing plasticity in speech perception comes from

studies with adults showing that the perception of novel phonemic contrasts can be improved by intensive auditory training. In a study in which young adult learners underwent 45 intensive training sessions, Logan et al. (1991) succeeded in improving the perception of the English /r/-/l/ contrast in Japanese-L1 speakers using a 'high variability phonetic training' approach. In this technique, learners are exposed to multiple tokens of minimal pairs of words containing the novel sounds in different syllable position produced by different talkers. These tokens are presented in a forced-choice identification task, with immediate feedback as to whether their choice was correct or incorrect. The aim of this approach is to create 'robust' novel phoneme categories by exposing L2 learners to acceptable variation within each category. The perceptual training was shown to be robust as it generalized to new words and new talkers (Lively et al. 1993b). Further studies replicated this result and showed that there was long-term retention of the learning over a period of three months at least (Lively et al. 1994; Bradlow et al. 1999). Perceptual training not only improved the perception of novel speech sounds but also led to improvements in the pronunciation of these sounds by the L2 learners that were also retained for at least three months following the completion of the training (e.g., Bradlow et al. 1997).

Perceptual training studies have mainly investigated consonant and vowel contrasts (e.g., Lambacher et al. 2002), with a strong focus on the English /r/-/l/ contrast. However, intensive auditory training has also been shown to be successful for suprasegmental features, leading to improvements in the perception of tone contrasts (e.g., Wang et al. 1999); again, improvement in the perception of these contrasts transferred to tone production, both for tokens heard during training and for novel tokens (Wang et al. 2003).

An important source of segmental information for speech perception that has not hitherto been fully exploited in intensive perceptual training is the information available via lipreading, which is available to language learners in traditional face-to-face language learning. For native listeners, visual cues to consonant and vowel identity are integrated with acoustic cues early in perceptual processing. This is shown by the McGurk effect (McGurk and MacDonald 1976), which is obtained when visual and acoustic cues to consonant identity are put in conflict (e.g., visual /g/ with auditory /b/). The resulting perception (typically /d/ in the example given) shows that the information from each modality is integrated before the sound is categorized by the listener. Visual cues add to the multiplicity of cues that are so crucial for making speech robust to environmental degradation, and are indeed particularly useful for native listeners in difficult listening conditions (e.g. Sumby and Pollack 1954).

It has been shown that, although infants are sensitive to visual information, as they prefer video presentations of speakers with congruent rather than incongruent audio

and visual channels (Kuhl and Meltzoff 1982), the use of lipreading cues develops with age. Indeed, children aged 6 to 10 years are less influenced by visual information in their judgments of consonant identity than are adult learners (Massaro et al. 1986; Sekiyama et al. 2003). Just as the sensitivity to certain acoustic cues increases within the first 10 years of life (e.g. Nittrouer and Miller 1997; Mayo and Turk 2004), sensitivity to visual cues also develops. The degree to which this visual influence develops appears to be language-dependent as McGurk effects show less visual influence in Japanese and Chinese speakers than English speakers (Sekiyama and Tohkura 1993, Sekiyama 1997, and Hardison 1999). The degree of visual influence may be related to the informativeness of visual cues in a particular language relative to auditory cues. This can depend on the number of visemes in a specific language (i.e. the number of 'visual categories' that are identifiable using lipreading alone), and also on whether the language is tonal as tone information is not marked visually. Interestingly, in a study looking at the development of visual influence in Japanese and English 6-year-old and adult listeners, it was shown that the level of visual influence was the same in 6-year-old English and Japanese listeners and that it increased significantly in English but not Japanese adults (Sekiyama et al. 2003). This suggests that, due to the relatively low informativeness of speechreading information in Japanese, Japanese listeners become strongly attuned to auditory rather than to visual information as a result of experience.

L2-learners may be therefore impeded in their use of visual cues for two reasons. First, if their first language is one in which visual influence is relatively low (e.g. Japanese or Chinese), the degree to which they will focus on segmental information available via the visual channel is likely to be low and they will need to be trained to attune more to the visual channel. Also, just as L2-learners lose sensitivity to acoustic cues to vowel or consonant contrasts that do not occur in their own language, it is probable that they lose sensitivity to even salient visual cues that are not relevant in their L1: they may be able to see the difference between two visible articulations but not be able to consistently associate each with the appropriate phoneme label. This was found to be the case in a comparison of the perception of English consonants and vowels presented with background noise in auditory and audiovisual conditions by Spanish-L1 speakers and native controls (Ortega-Llebaria, Faulkner and Hazan 2001). Consonant confusions that were language-dependent, i.e. which resulted from differences in the phoneme inventories of Spanish and English, were not reduced by the addition of visual cues, whereas confusions that were common to both listener groups and were due to acoustic-phonetic similarities between categories did show improvements. A wider-ranging study of concordant and discordant audiovisual syllables with listeners from four different L1 backgrounds showed that the influence of the visual cue was "dependent upon its information value, the intelligibility of the auditory cue and the

assessment of similarity between the two cues" and was also affected by linguistic experience (Hardison 1998). Even if L2 learners are not highly sensitive to visual cues, an important issue is whether this sensitivity can be increased by intensive training. Improving the use of visual information to consonant identity is likely to be effective in improving speech perception in L2 learners, as it would increase the multiplicity of phonetic information available to these listeners.

Few studies to date have evaluated the relative effectiveness of intensive auditory and audiovisual training on the acquisition of novel L2 contrasts. In a study on the training of the American English /r/-/l/ contrast with Japanese and Korean learners of English, audiovisual training was more effective than purely auditory training in improving the identification of the contrast for both learner populations. Improvements in perception also led to improvements in native listener ratings of the learners' pronunciation both for the auditory-trained and audiovisual-trained learners (Hardison 2003). Another approach to audiovisual training has been to use three-dimensional computer-animated talking heads, such as Baldi (Massaro 1999) that use accurate articulatory information. These are potentially powerful tools for language learning as training materials can be easily generated and as they give greater potential for 'enhancing' articulatory information, by making the skin transparent so that articulatory gestures within the oral tract can be seen, by slowing down articulatory gestures or highlighting usually invisible cues such as vocal fold vibration (Massaro 1998). Such talking heads have already successfully been used for language learning by children with hearing loss (Massaro and Light 2004). In the first study investigating the usefulness of a talking head for training the /l/-/r/ contrast with Japanese learners of English (Massaro and Light 2003), two training conditions were compared with a) the 'standard' talking head and b) the same talking head with 'transparent' skin and visible articulatory gestures. Training significantly improved the identification and production of /l/ and /r/ but the visible articulation condition did not lead to a great improvement. This study does not permit us to compare the relative effectiveness of natural and synthetic visual cues.

2. Effect of perceptual training on speech production

A number of previous studies have shown that perceptual improvements resulting from intensive training transfer to the pronunciation of the trained sounds. The aim of this experiment is to evaluate whether the effect could be replicated here, and whether audiovisual training of a contrast (i.e. seeing the articulatory gestures of the talker) would lead to a greater improvement in pronunciation, even if did not lead to a greater improvement in perception for the less visually-salient /l/-/r/ contrast.

2.1 Speakers and Listeners

Twenty-five Japanese students were recorded before and after their course of training. Ten had completed the training study in the 'A training' condition, ten in the 'AV natural' condition and 5 in the 'AV synthetic' condition.

The listeners were native speakers of British English tested in London.

2.2 Speech materials and recording conditions

At the time of the pre- and post-test, the Japanese trainees were asked to read a list of 25 minimal pairs of words which included the sounds /l/ and /r/ in a variety of phonetic environments and positions. The audio recordings were made in a soundproof room using a Sony TCD-D10 digital audio recorder with a ECM-959DT microphone. The sounds /l/ and /r/ appeared in initial and medial positions, in singleton and clusters. The minimal pairs were: lake-rake, long-wrong, miller-mirror, blight-bright, complies-comprise. A total of 20 pre/post tokens per trainee were therefore included, yielding a total of 500 tokens for 25 trainees. These were normalized for level by equating the peak amplitudes.

2.3 Experimental tasks

Two independent perceptual evaluation tests were carried out in which native speakers of British English were asked to judge the tokens produced by the learners. The tests included a minimal-pair identification task and a quality rating task. For each test, the listeners evaluated the productions of all learners. They were asked to focus on the /l/-/r/ realizations in the word rather than on the correct pronunciation of the word itself.

The experiment was run on a laptop and items were presented to both ears at a comfortable listening level via headphones. The identification test took about 45 minutes, the rating task about 55 minutes.

2.4 Minimal-pair identification task

Tokens were presented in a two-alternative forced choice task. In each trial, the minimal pair appeared in writing on two buttons on the screen, and listeners heard the token of a learner and indicated their choice of the word by pressing the button. The order of the items was randomized across trials.

2.5 Consonant rating task

Listeners were asked to judge the realizations of /l/ and /r/. The intended word appeared on the screen, followed by the auditory prompt. Listeners rated the /l/ or /r/ in the word on a scale from 1-7 (1 = bad, 7 = excellent). The order of presentation was

randomized across trials.

2.6 Results

2.6.1 Minimal pair identification test

For each of the 12 listeners, mean identification scores were calculated for /r/ and /l/ for each of the 25 learners. The percentage of /l/ and /r/ productions that were correctly identified by native listeners is shown in Table 1. It can be seen that the correct identification of /r/ tokens produced by the learners is generally much better than the identification of /l/ tokens, which is consistent with previous findings on the production of the /r/-/l/ contrast by Japanese learners (Bradlow et al. 1997).

Table 1: Percentage of correct consonant identification for /l/-/r/ in words produced prior and following training for learners in the Auditory, AV (natural face) and AV synthetic face training groups.

Training Mode		/l/		/r/		Mean intelligibility	
		Pre	Post	Pre	Post	Pre	Post
Auditory	Mean	54.8	59.8	86.0	83.5	70.4	71.7
N = 120	Std. Dev.	*23.3*	*28.5*	*17.7*	*22.7*	*14.7*	*12.7*
AV (natural face)	Mean	60.7	67.3	76.2	82.7	68.4	75.0
N = 120	Std. Dev.	*30.2*	*24.6*	*24.5*	*24.5*	*17.4*	*14.0*
AV (synthetic face)	Mean	60.7	65.3	96.0	96.0	78.3	80.7
N = 60	Std. Dev.	*34.1*	*25.2*	*8.1*	*8.9*	*16.3*	*14.0*

The data suggest that learners in the 'AV synthetic' group had a significantly better pronunciation prior to training than other groups, as /r/ productions for this group were almost perfectly identified (mean of 96% correct identification). Because of the ceiling effect for /r/ identification, we therefore removed the data for the 'AV synthetic' learners before subsequent data analysis and compared the 'A' and 'AV natural' groups only. A repeated-measures ANOVA was carried out on the identification scores for /r/ and /l/ obtained before and after training. The effect of time of testing was significant, showing that identification scores improved significantly post-training [$F(1,238) = 20.0$; $p < 0.001$]. Evaluations of the effect of training mode were carried out on the scores reflecting the difference in identification scores pre-post training for /r/ and /l/ to normalize for differences in pre-training scores. The increase in identification scores was significantly greater for the tokens produced by the 'AV natural' training learners than for those for the 'A training' learners [$F(1,238) = 9.271$; $p < 0.005$]. The within-group effect of consonant was not significant.

The following analysis focused on individual differences among learners in terms of the effect of perceptual training on production. The change in mean identification scores from pre- to post-training tokens produced by individual learners ranged from −11% to +20%. A univariate ANOVA showed that this effect was significant [$F(19,220) = 8.538$; $p < 0.001$], suggesting that some learners improved their pronunciation more significantly than others.

2.6.2 Consonant rating task

For each of the 12 listeners, consonant rating scores were calculated for the /r/ and /l/ sounds produced by each of the 20 learners from the A and AV (natural face) groups before and after training (See Table 2). The data obtained for the rating task closely mirror those obtained for the consonant identification task. Repeated-measures ANOVAs revealed that ratings were higher for the post-training tokens than the pre-training tokens [$F(1,238) = 33.6$; $p < .001$]. The effect of training mode was assessed from the difference in ratings before and after training. The effect of training mode was not significant, but there was a significant consonant by training interaction [$F(1,238) = 6.48$; $p < 0.02$]. Post-hoc tests showed that /r/ rating improved more for the AV training group than the A training groups but that /l/ ratings did not. Figure 1 summarizes the pre/post identification scores and ratings for each training group.

Table 2: Consonant rating on a scale of 1 to 7 (1 = bad, 7 = excellent) for words produced prior and following training by learners in the Auditory, AV (natural face) and AV synthetic face training groups.

Training		/l/		/r/		Mean rating	
		Pre	Post	Pre	Post	Pre	Post
Auditory	Mean	3.8	4.1	5.0	5.0	4.4	4.6
	Std. Dev.	*1.1*	*1.2*	*0.9*	*1.1*	*0.8*	*0.7*
AV (natural face)	Mean	4.2	4.4	4.5	4.9	4.4	4.6
	Std. Dev.	*1.3*	*1.1*	*1.1*	*1.1*	*0.9*	*0.8*
AV (synthetic face)	Mean	4.0	4.3	5.5	5.2	4.7	4.8
	Std. Dev.	*1.4*	*1.1*	*0.8*	*0.9*	*0.9*	*0.9*

3. Discussion

The finding that perceptual training results in improvements in the pronunciation of the trained consonants in second language learners is consistent with many other studies of perceptual training. A key finding of our study was that, even for the less visual-

ly-distinctive contrast, the use of audiovisual stimuli with a natural face in intensive perceptual training led to a greater improvement in the production of the difficult consonants compared to learners trained with auditory stimuli, even though there was no difference between the training groups in terms of the effect of training on the perception of the /l/-/r/ contrast. Exposure to the visible articulatory gestures involved in the production of /l/ and /r/ therefore seems to have been effective, even without specific pronunciation training. This finding needs to be verified with a wider range of phonemic contrasts and speakers. The planned comparison between the effect of natural and synthetic articulatory gesture information was not achieved due to the small number of learners in the AV synthetic group and imbalance in pre-training production performance.

The fact that intensive perceptual training can result in improvements in the pronunciation of the sounds being trained has important practical implications for computer-assisted language learning. Indeed, computer-based perceptual training programmes are more reliable than computer-based pronunciation training programmes, which need to provide accurate automatic ratings of the learner's productions. Contrary to what might be expected from speech perception studies with native speakers, audiovisual presentation of stimuli may not automatically produce gains in perception relative to auditory stimuli, especially for speech contrasts that may be visible but not highly salient. However, further investigations need to be carried out to investigate whether training which further enhances visual cues may be more successful in attracting language learners' attention to this useful source of segmental information.

4. Acknowledgements

I would like to thank Dr. Valerie Hazan, Anke Sennema, Andrew Faulkner (University College London), and the UCL Language Center for their substantial help in organizing the testing of Japanese students.

References

Best, C., 1995. A direct realist view of cross-language speech perception, in: Strange, W. (ed.), *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, York Press,. Baltimore, pp. 171-204.

Best, C., McRoberts, G., and Goodell, E., 2001. *Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system.* Journal of the Acoustical Society of America, 109, 775-794.

Bradlow, A., Pisoni, D., Akahane-Yamada, R. and Tohkura, Y., 1997. Training Japanese listeners to identify English /r/ and /l/: IV, some effects of perceptual learning on speech production. Journal of the Acoustical Society of America, 101, 2299-2310.

Cole, R., 1999. *Tools for research and education in speech science*, Proc. ICPhS.

Demorest M.E., Bernstein L.E., DeHaven G.P., 1996. Generalizability of speechreading performance on nonsense syllables, words, and sentences: Subjects with normal hearing, J. Speech Hear. Res., 39, 697-713.

Flege, J.E., 1995. Second-language speech learning: theory, findings, and problems, in: Strange, W. (Ed.), *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, York Press,. Baltimore, pp. 229-273.

Flege, J.E., 1998. *Second-language learning: the role of subject and phonetic variables*, Proc. Still ESCA workshop, Stockholm, May 1998, 1-8.

Flege, J.E., 1999. Age of learning and second-language speech. In: Birdsong, D.P. (Ed.), *Second Language Acquisition and the Critical Period Hypothesis, Lawrence* Erlbaum, Hillsdale, NJ, pp. 101-132.

Fowler, C., 1986. *An event approach to the study of speech perception from a direct-realist perspective*, J. Phon., 14, 3-28.

Gagne J.-P., Rochette A.-J., Charest M., 2002. *Auditory, visual and audiovisual clear speech, Speech* Com., 37, 213-230.

Grant, K.W., Seitz, P.F., 1998. *Measures of auditory — visual integration in nonsense syllables and sentences*, J. Acoust. Soc. Am., 104, 2438-2450.

Guion, S.G., Flege, J.E., Ahahane-Yamada, R. & Pruitt, J.C., 2000. *An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants*, J. Acoust. Soc. Am., 107, 2711-2725.

Hardison, D., 1999. *Bimodal speech perception by native and nonnative speakers of English: Factors influencing the McGurk effect, Language Learning*, 49, 213-283.

Hardison, D., 2003. *Acquisition of second-language speech: Effects of visual cues, context and talker variability, Applied Psycholinguistics*, 24, 495-522.

Iba, M., 2004. *The effectiveness of visual cues* in L2 perception, The Journal of the Institute for Language and Culture, 8, 43-56.

Iverson, P., Kuhl, P.K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., Siebert, C., 2003. *A perceptual interference account of acquisition difficulties for non-native phonemes*, Cognition, 87, B47-B57.

Kuhl P.K., 1993. *Early linguistic experience and phonetic perception: implications for theories of developmental speech perception.* J. Phon., 21, 125-139.

Kuhl, P.K. & Meltzoff, A.N., 1982. *The bimodal perception of speech in infancy*, Science, 218, 1138-1141.

Lambacher, S., Martens, W., Kakeki, K., 2002. *The influence of identification training on identification and production of the American English mid and low vowels by native speakers of Japanese. Proc. 7th Int. Conf. Spoken Lang. Proc., Denver, 245-248.*

Lenneberg, E.H., 1967. *Biological Foundations of Language.* John Wiley and Sons Inc, New York.

Logan, J.S., Lively, S.E., Pisoni, D.B., 1991. Training Japanese listeners to identify English /r/ and /l/, J. Acoust. Soc. Am., 89, 874-886.

Lively, S., Logan J., Pisoni, D., 1993. Training Japanese listeners to identify English /r/ and /l/. II: *The role of phonetic environments and talker variability in learning new perceptual categories*, J. Acoust. Soc. Am., 94, 1242-1255.

Lively, S.E., Logan, J.S., Pisoni, D.B., 1993b. *Training Japanese listeners to identify English /r/ and /l/. III: long-term retention of new phonetic categories*, J. Acoust. Soc. Am., 94, 1242-1255.

Marassa L.K., Lansing C.R., 1995. *Visual word recognition in 2 facial motion conditions — full face versus lips-plus-mandible*, J. Speech Hear. Res., 38, 1387-1394.

Massaro, D.W., 1998. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle.*, MIT Press, Cambridge, MA.

Massaro, D.W., Cole, R. 2000. From *"Speech is special" to talking heads in language learning*, Proc. INSTIL, Dundee, Scotland, 153-161.

Massaro, D.W., Cohen, M.M., Daniel, S., Cole, R.A., 1999. Developing and evaluating conversational agents, in: P.A. Hancock (Ed.), Human Factors and Ergonomics: *Perceptual and Cognitive Principles.* (Handbook of Perception & Cognition, 2nd Edition), Academic Press, San Diego, pp. 173-194.

Massaro, D.W., Light, J., 2003. Read My Tongue Movements: *Bimodal Learning To Perceive And Produce Non-Native Speech /r/ and /l/.* Proc. Eurospeech 2003, Geneva, Switzerland, 2249-2252.

Massaro, D.W., Light, J., 2004. *Using visible speech to train perception and production of speech for individuals with hearing loss.* J. Speech Lang. Hear. Res., 47, 304-320.

Massaro, D.W., Thompson, L.A., Barron, B., Laren, E., 1986. *Developmental changes in visual and auditory contribution to speech perception*, J. Exp. Child Psychol., 41, 93-113.

Mayo, L.H., Florentine, M., Buus, S., 1997. *Age of second-language acquisition and perception of speech in noise, J. Speech*, Lang. Hear. Res., 40, 686-693.

Mayo, C., Turk, A., 2004. Adult-child differences in acoustic cue weighting are influenced by segmental context: Children are not always perceptually biased toward transitions, J. Acoust. Soc. Am., 115, 3184-3194.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices, Nature, 264, 746-748.

Nittrouer, S., Miller, M.E., 1997. Predicting developmental shifts in perceptual weighting schemes, J. Acoust. Soc. Am., 101, 2253-2266.

Ortega-Llebaria, M., Faulkner, A., Hazan, V., 2001. Auditory-visual L2 speech perception: *Effects of visual cues and acoustic-phonetic context for Spanish learners of English.* Proc. AVSP-2001, 149-154.

Ouni, S. Massaro, D.W., Cohen, M.M., Young, K., Jesse, A., 2003. *Internationalization of a talking head*, Proc. ICPhS, Barcelona, Spain, August 2003, 2569-2572.

Sekiyama, K., Tohkura, Y., Umeda, M., 1996. *A few factors which affect the degree of incorporating lip-read information into speech perception*, Proc. ICSLP1996, Denver, USA, 1481-1484.

Sekiyama, K., Tohkura, Y., 1993. *Inter-language differences in the influence of visual cues in speech perception'*, J. Phon., 21, 427-444.

Sekiyama, K., 1997. *Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects*, Perc. Psychophys., 59, 73-80.

Sekiyama, K., Burnham, D., Tam, H., Erdener, D., 2003. *Auditory-visual speech perception development in Japanese and English speakers*. Proc. AVSP 2003, St Jorioz, France, September 2003, 43-47.

Sennema, A., Hazan, V., Faulkner, A., 2003. *The role of visual cues in L2 consonant perception, Proc. 15th ICPhS, Barcelona, Spain, August 2003, 135-138.*

Siciliano, C., Faulkner, A., Williams, G., 2003. *Lipreadability of a synthetic talking face in normal hearing and hearing-impaired listeners*, Proc. *AVSP 2003*, St Jorioz, France, September 2003, 205-208.

Sumby, W.H. and Pollack, I., 1954. *Visual contribution to speech intelligibility in noise'*, J. Acoust. Soc. Am., 26, 212-215.

Summerfield, Q., 1983. Audio-visual speech perception, lipreading and artificial stimulation, in: M.E. Lutman & M.P. Haggard (eds), *Hearing Science and Hearing Disorders*, Academic Press, London, pp. 131-182.

Takata, Y., Nabelek, A.K., 1990. *English consonant recognition in noise and in reverberation by Japanese and American listeners*, J. Acoust. Soc. Am., 88, 663-666.

Tyler, M.D., 2001. Resource consumption as a function of topic knowledge in nonnative and native comprehension, *Language Learning*, 51, 257-280.

Wang, Y., Spence, M.M., Jongman, A., Sereno, J.A., 1999. *Training American listeners to perceive Mandarin tones*, J. Acoust. Soc. Am., 106, 3649-3658.

Wang, Y., Jongman, A, Sereno, J.A., 2003. *Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training*, J. Acoust. Soc. Am., 113, 1033-1043.

Werker, J.F., Tees, R.C., 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life, *Infant Behavior Development*, 7, 47-63.

Yamada, R.A. 1995. Age and acquisition of second language speech sounds: perception of American English /r/ and /l/ by native speakers of Japanese in: Strange, W. (ed.), Speech Perception and Linguistic Experience: *Theoretical and Methodological Issues*, York Press,. Baltimore, pp. 305-320.