ENHANCING ACOUSTIC CUES TO AID L2 SPEECH PERCEPTION

Marta Ortega and Valerie Hazan
Dept of Phonetics & Linguistics, UCL

ABSTRACT

Hazan and Simpson [3] showed that enhancement techniques that selectively amplify regions of the speech signal containing information to consonant identity increased consonant intelligibility for VCV utterances presented in noise for both native and non-native speakers. Here, we explored the relationship between the effect of enhancement on perception of speech in quiet by non-native listeners and ease of perception of L2 phoneme categories, predicted on the basis of the degree of assimilation of L2 contrasts to L1 categories. The perception of three English contrasts (/r/-/l/, /w/-/r/, /w/-/j/ for Japanese listeners and /d/-/ð/, /t/-/d/, /ð/-/t/ for Spanish listeners) was assessed by means of syllable matching, identification and discrimination tasks. Post-hoc analyses of the error rates in identification tasks indicated that our enhancement techniques improved the perception of those English contrasts that were assimilated to a single L1 category in natural speech. Thus, enhancements led to increased perception in speech in quiet without previous training.

1. INTRODUCTION

It is well known that second language learners have particular difficulty in perceiving phonemic contrasts that do not occur in their native language. According to Best's Perceptual Assimilation Model [1], ease of perception of L2 contrasts may be predicted on the basis of the degree of assimilation of L2 contrasts to L1 categories: two L2 sounds that are assimilated to a single L2 category are predicted to be very difficult to discriminate whereas L2 sounds which are assimilated to different L1 categories may differ in discriminability.

Many studies [2, 4, 6, 7] have explored the effect of auditory training in improving the perception of non-native contrasts. The aim of such training methods is to focus the listener's attention on phonetically-relevant dimensions within the speech signal. Some have had some success not only in showing improvements in perception but also generalisation to other stimuli or speakers and maintenance of the effect over a period of months, but training is time-consuming and effortful [2, 4]. Another approach that has been shown to be successful in improving consonant intelligibility by directing the listener's attention to phonetically-relevant parts of the signal is cue-enhancement (e.g., [3]). In this technique, regions of the speech signal containing information to consonant identity are selectively amplified. Nonsense VCV utterances that were enhanced in this way have been shown to be more intelligible in noise both for native and non-native listeners [3]. The aim of this study was (1) to investigate whether such a technique would also be successful in increasing the intelligibility of L2 consonant contrasts presented in quiet and (2) whether any effect of enhancement would be dependent on the degree of assimilation of L2 consonants to L1 categories. This study consisted of two phases; first, the degree of assimilation of L2 consonants to L1 categories was evaluated for three consonant contrasts using a consonant assimilation task. Then, L2 listeners' intelligibility of natural and enhanced versions of the consonants was evaluated using identification and discrimination tasks.

2. CONSONANT ASSIMILATION TASK

2.1. Test material

Two sets of English phonemic contrasts were used. The first set, which included /t-d/, /t-ð/, /ð-d/ was used to test the perception of Spanish speakers of English. The second set, /r-l/, /r-w/, /w-j/, examined consonant perception by Japanese speakers of English. Target syllables consisted of CV stimuli in three vocalic contexts: for Spanish speakers, /t/, /d/, /ð/ were presented in the context of the vowels /ɑ/, /i/, and /o/, and for Japanese /r/, /l/, /w/, /j/ were presented with /ɑ/, /o/, and /e/.

2.2 Stimuli

Stimuli were recorded by four native speakers of South Eastern British English accent (2 male and 2 female), aged between 25 to 35. In a sound-treated room, they read aloud several repetitions of the target syllables and words within the carrying sentence "Say _ please" and recordings were made onto a DAT tape. The recorded material was then digitised at a 20 kHz sampling rate with 16-bit amplitude quantisation. Two experimenters selected the clearest repetitions to be used in the master tapes.

2.3. Experimental task

Syllable tests explored the assimilation patterns of the tested English phonemes to L1 categories (Spanish or Japanese). They contained four repetitions of each target syllable per speaker. Repetitions were randomized, grouped in sets of 20, and played every 2.5 seconds with a 4 second break after each block of 20. Listeners were instructed that the English speakers were learning Japanese (or Spanish) and had to identify the Japanese (or Spanish) consonant that the English speakers intended to say.

2.4. Listeners

Thirty ESL speakers (12 Japanese with mean age of 27 yrs, s.d. 7.31), 18 Spanish with mean age of 24.50 yrs, s.d. 9.26) who were attending a two week summer school on English phonetics in the UK participated in the experiment. The mean age at which subjects started learning English was 10.11 yrs (s.d. 2.83) for Japanese subjects and 12.08 yrs (s.d. 0.79) for the Spanish subjects. Listeners were in most cases monolingual, and none spent more than 12 months living in an English speaking country, although they had been studying English in their own countries for at least 6 years.

2.5. Test procedure

Listeners were tested in groups of 4 to 10. They first completed a questionnaire that gathered information about their language background. The consonant assimilation task was then presented.

2.6. Results

The matrices in Table 1 show the listeners' assimilation patterns between the English sounds and their L1 categories. Spanish speakers mainly assimilated English /t/ to Spanish /t/, English /d/ to Spanish /d/ and /t/, and English /ð/ to Spanish /d/ and /θ/. Japanese listeners mainly assimilated English /l/ and /r/ to Japanese /r/, English /j/ to Japanese /j/, and English /w/ to Japanese /w/.

Spanish

English	t	d	θ	tʃ	oth
t	9	6	0	1	0
d	3	6	0	0	1
ð	2	8	1	0	1

Japanese

	j	r	w	oth
j	9	0	0	2
l	0	10	0	0
r	0	9	2	1
w	0	2	9	8

Table 1. Matrices showing identification percentages

2.7. Discussion

The assimilation patterns obtained in the syllable tests can be interpreted as One Category or Two Category contrasts as described in the Perceptual Assimilation Model (PAM; [1]), if assimilation is defined as a 'high percentage' of identification of an L2 sound with an L1 sound. The patterns of assimilation that were obtained are summarised in Figure 1.

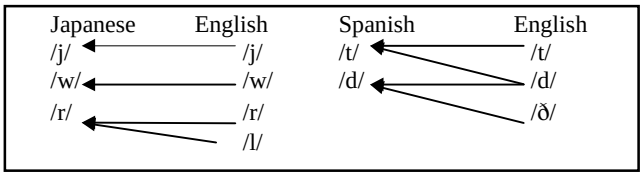

Figure 1. Japanese and Spanish assimilation patterns to English

These assimilation patterns predict that Japanese ESL speakers will perceive the English /r-l/ contrast as a One-Category contrast and the English contrasts /r/-/w/ and /w/-/j/ as Two-Category as Japanese ESL speakers assimilated each of these sounds to a different Japanese phoneme. Spanish speakers perceived the English contrasts /d̪/ð/ and /t/-/d/ as One-category contrasts, and /t/-/ð/ as a Two-Category contrast. Moreover, PAM relates One- and Two-Category contrasts with ease of perception. Two-Category contrasts are easy to perceive, and therefore, good identification and discrimination scores are predicted. One-Category contrasts can vary in their level of perceptual difficulty, and as a whole, are more difficult to perceive than Two-Category contrasts. Thus, the results of the syllable test should predict the perceptual difficulty that our L2 subjects encountered with the English contrasts presented in the identification and discrimination tasks.

3. IDENTIFICATION AND DISCRIMINATION TASKS

Identification and discrimination tasks were used to evaluate the intelligibility of minimal pairs of words containing the phonemic contrasts described above. Stimuli were presented both in natural and enhanced conditions for both perceptual tasks.

3.1. Test material

The target words used for Spanish and Japanese listeners are listed in Table 2. Each phonemic contrast was tested in three different vocalic contexts.

	/d-ð/		/t-d/		/t-ð//	
/ɛ/	dense	thence	tense	dense	tense	thence
/əʊ/	doze	those	toes	doze	toes	those
/ɑɪ/	die	thy	tie	die	tie	thy
	/r-l/		**/r-w/**		**/w-j/**	
/o/	raw	law	raw	war	war	yaw
/æ/	rack	lack	rack	wack	wack	yak
/ɪ/	rip	lip	rip	whip	whip	yip

Table 2. English contrasts presented to Spanish speakers (top) and Japanese speakers (bottom).

3.2. Methodology

In order to create the enhanced stimuli to be used in the identification and discrimination tasks, speech files were annotated using a waveform-editing tool to mark the constriction/occlusion consonantal regions and the first five cycles of the vowel following the consonant. For the vowel onset region, the reduced amplitude as the consonant constriction/occlusion was released was counteracted by progressively amplifying the initial five cycles of the second vowel by 8 dB (first cycle) to 4 dB (fifth cycle). For /d/ and /t/, the burst transient was amplified by 12 dB and aspiration regions in stops by 6 dB. For /ð/, the friction region was amplified by 6 dB and for /w, j, l, r/ the glide regions by 9 dB. The amplification was applied digitally by scaling the regions' sample values. To avoid waveform discontinuities at region boundaries, 5 ms raised-cosine ramps were used to blend adjoining sections together. Enhanced ([235_sp_enh.wav], [235_jap_enh.wav]) and natural stimuli ([235_sp_nat.wav],[235_jap_nat.wav]) were set at a constant rms level of 23 dB.

3.3. Experimental tasks

For the identification task, one repetition of the natural and enhanced versions of the words listed in Table 2 produced by four speakers were included, giving a total of 144 tokens. Order of presentation of stimuli was fully randomised. Inter-stimulus interval was 1.5 seconds, and a longer break was provided after each block of 20 items. Listeners were instructed to circle, on a printed answer sheet, which of the two words of the minimal pair printed on each line had been heard.

An AXB procedure was used for the discrimination task. In this task, each contrast was only presented in two vocalic contexts /əʊ/ and /ɑɪ/ for Spanish listeners, and /æ/ and /ɪ/ for Japanese listeners. A single repetition of the natural and enhanced versions produced by four speakers was included, yielding a total of 192 tokens. The inter-stimulus time was 0.5-0.5-2.0 seconds with a longer break after each block of 20 items. Listeners were instructed to circle 'A' on the response sheet if the middle word was the same as the first and 'B' if it was the same as the third word heard.

3.4. Subjects
As above.

3.5. Results
In the identification tasks, in general, L2 listeners made fewer mistakes in their perception of enhanced stimuli than of natural stimuli. A within-subject three-factorial ANOVA was performed for each task and language group. The three factors were condition (enhanced versus natural stimuli), vowel context, and c

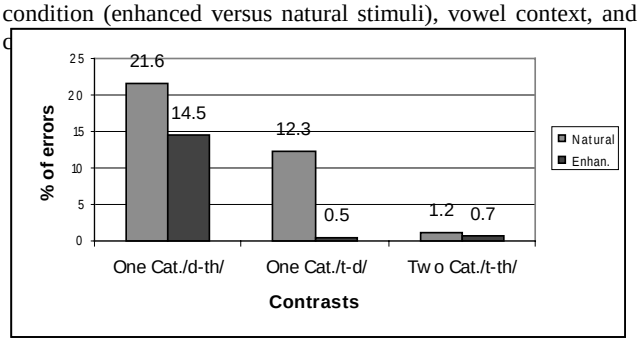

Figure 2. Performance of Spanish speakers in the identification task

In the ANOVA for the identification task and Spanish group (see Figure 2), a significant effect of condition [F (1,17)=32.329; p<0.0001], contrast [F (2,34)=14.412; p<0.0001], context [F (2,34)=6.446; p=0.004], and their interactions was found. Multiple pairwise comparisons were performed for the contrast factor for natural and enhanced stimuli separately, since subjects' performance in each contrast changed with stimuli type. Only the Two-Category /t/-/ð/ contrast differed significantly from the other two in the natural stimuli (p<.0001 and p<.0001), and the One-Category /d/-/ð/ contrast from the One-Category /t/-/d/ and the Two-Category /t/-/ð/ in the enhanced stimuli (p=.010 and p=.014). The interaction between contrast and condition [F (2,34)=11.009; p<0.0001] showed that listeners made fewer errors in enhanced than in natural stimuli for the /d/-/ð/ and /t/-/d/ contrasts. In the Two-Category /t/-/ð/ contrast, however, there was no improvement since there were virtually no errors.

For Japanese listeners (See Figure 3), in the ANOVA performed on the identification data, only the effect of contrast [F (2,22)=18.415; p<0.0001] and the interactions of 'condition x context' [F (2,22)=6.673; p=0.005], and 'condition x context x contrast' [F (4,44)=2.622; p=0.047] were significant.

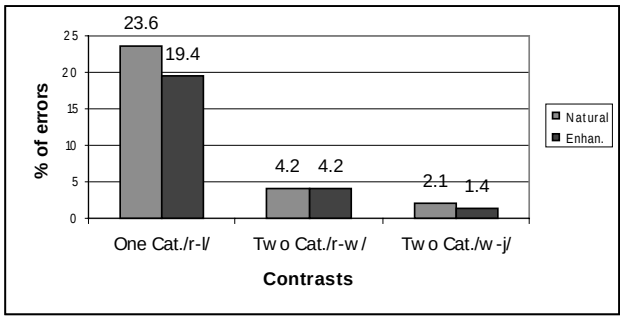

Figure 3. Performance of Japanese speakers in identification task.

Multiple pairwise comparisons for natural and enhanced stimuli, showed that the One-Category /r-l/ contrast differed significantly from the two Two-Category contrasts (p=.021 and p=.005 in enhanced stimuli, and p=.008 and p=.003 in natural stimuli). The significant interactions indicated that for the contexts /æ/ and /o/ and for the most difficult contrast, i.e. /r-l/, Japanese listeners tended to make fewer errors in enhanced than in natural stimuli. Japanese listeners made very few errors for the Two-Category contrasts in both conditions, and consequently, enhancement did not trigger a substantial reduction of errors. As for the /ɪ/ context, Japanese listeners made more mistakes in enhanced than in natural stimuli in the One-Category contrast. Once words containing /ɪ/ were eliminated from the ANOVA, both contrast and condition became significant main effects [F (2,22)=18.662; p<0.0001 and F (1,11)=9.842; p=0.009].

Japanese and Spanish speakers were highly accurate in the discrimination task (mean percentage correct: 94.95% for natural stimuli, 95.28% for enhanced stimuli). ANOVAs showed the effect of 'contrast' to be the only significant factor common to both language groups. The effect of context was significant in Spanish and the interaction of 'contrast x context' for Japanese. Multiple pairwise comparisons for the contrast were performed separately for natural and enhanced stimuli. In natural stimuli, the Two-Category /t/-/ð/ contrast differed significantly from the other two in Spanish (p=.011 and p=.014), and the One-Category /r/-/l/ from the Two-Category /w/-/j/ in Japanese (p=.026). In enhanced stimuli, /r/-/l/ differed significantly from /w/-/j/ (p=.027).

In summary, for identification tasks, the effects of condition (natural vs. enhanced) and contrast were significant for both language groups once the context /ɪ/ was eliminated from the Japanese data. 'Contrast x condition x context' was the significant interaction common to both language groups. It showed that subjects made fewer errors in enhanced than in natural stimuli for the One-Category contrasts in Spanish and in Japanese. These patterns repeated in all vocalic contexts except for /ɪ/ in Japanese data. In discrimination tasks, the effect of contrast was the only common significant factor for Japanese and Spanish. For natural stimuli, the /t/-/ð/ contrast differed significantly from the other two in Spanish and the /r/-/l/ from /w/-/j/ in Japanese.

Previous studies [3] showed that the extent of the enhancement effect varied across listeners and across speakers. Here, the effect of enhancement was consistent for a majority of listeners: for the Spanish group, no listeners obtained lower scores for the enhanced condition, and for the Japanese, only three out of 12 did (i.e. two listeners obtained –1% and one –2%). Increases in intelligibility ranged from 0 to 16.89% in the Spanish group, and up to 7% in Japanese.

As for speakers, all four were assigned fewer errors in the enhanced stimuli by Spanish listeners, and two by Japanese listeners. Paired t-tests showed that these differences were significant in three of the speakers (p=.019, .029, .045) for the Spanish group, in one speaker (p=.007) for the Japanese group.

5. DISCUSSION
Results confirmed predictions as both Japanese and Spanish

listeners experienced different levels of perceptual difficulty for the tested English contrasts. In the natural condition, Spanish listeners made more errors for the One-Category /t/-/d/ and /ɗ/-/ð/ English contrasts, and fewer for the Two-Category /t/-/ð/ contrast. Japanese listeners found the One-Category /r/-/l/ English contrast difficult and the Two-Category /w/-/j/ and /w/-/r/ contrasts to be easy to perceive as shown by error rates obtained. These results assign the same level of perceptual difficulty to the English contrasts than those predicted by the consonant assimilation test. Therefore, they provide independent evidence for PAM's claim [1] that ease of perception of L2 contrasts by L2 beginners can be predicted by the assimilation patterns that these speakers establish between L2 and L1 sounds.

Pairwise comparisons of contrasts in enhanced stimuli also indicated that the three contrasts were grouped into two levels of perceptual difficulty. Japanese listeners perceived the One-Category /r/-/l/ contrast as difficult and the Two-Category /r/-/w/, /w/-/j/ contrasts as easy to perceive, while Spanish speakers categorised One-Category /ɗ/-/ð/ as difficult, and /t/-/d/ (One-Category) and /t/-/ð/ (Two-category) as easy. Thus, for Spanish data, /t/-/d/ changed 'membership' from 'as difficult to perceive as / /-/ð/' to 'as easy as /t/-/ð/'. This change of membership interpreted together with the relationship between assimilation patterns and level of perceptual difficulty illustrates some of the effects of enhancement on L2 perception. Enhancement helped Spanish speakers to hear English /t/ and /d/ as two different phonemes as they did in the /t/-/ð/ contrast.

The effect of enhancement in L2 perception is further illustrated by the significant main effect of condition and the interaction 'condition x contrast x context' in identification tasks[1] in both language groups. While the effect of condition demonstrated that enhancement had an effect in improving L2 perception, at least in the identification tasks, the interaction showed that this effect depended on type of contrast and context. As for context, improvement took place in all except for the /ɪ/ context in Japanese. With regards to contrast, enhancement significantly reduced error rate in the /r/-/l/, /t/-/d/, and /d/-/ð/ contrasts, which were classified as One-Category contrasts in natural stimuli. Thus, enhancement techniques embedded in word minimal pairs improved the perception of phonemic mergers, or One-Category contrasts, in both L1 groups. These techniques helped L2 listeners to perceive the acoustic cues that did signal a difference in meaning in L2 while ignoring those cues that did not, even though the distinctive L2 cues were not used or used differently in their L1 system.

As recent literature on L2 perception points out (e.g, [6]), the above changes in L2 perception can be explained as changes on selective attention during categorisation and perceptual learning (e.g., [5]). Subjects assign different relevance to each dimension of the object when classifying objects into different categories. Since a successful classification requires processing only those dimensions that differentiate categories, subjects should focus their attention to these relevant dimensions. Therefore, learning to classify a set of objects into two different categorisation systems requires that subjects focus their attention to the relevant dimensions for each categorisation system. For example, our Spanish subjects categorised 12.27% of English /d/ as Spanish /t/ in natural stimuli, while this error percentage reduced to 0.46% in enhanced stimuli. These differences in performance can be explained if Spanish subjects paid more attention to the pre-voiced / non-pre-voiced dimension in natural stimuli while they focused on the aspiration cue in enhanced stimuli during /t/-/d/ identification. Thus, within this theory of categorisation, enhancement can be thought as the trigger that switches focus of attention to the relevant L2 acoustic cues.

This experiment differed from the training studies mentioned above with regards to the technique used in redirecting the attention of L2 listeners to different acoustic cues, amount of improvement obtained, and time spent to achieve improvement. Using training techniques on the perception of the /r-l/ contrast with Japanese ESL subjects, Lively [6], for example, obtained around 7% improvement in word initial singletons after 3 weeks of training. Bradlow's subjects [2] improved their performance for word initial /l/ from 45% to 85% approximately while they experienced no gain in /r/ after 45 training sessions. In this study, 4.2% improvement was obtained in word initial singletons in one 40 minute session and without any training. The immediate perceptual improvement makes enhancement a potentially useful technique for auditory training and for speech technology applications for L2 users. Improvement obtained in training techniques has been shown to generalize to unknown stimuli, to last for at least three months [6], and ameliorate production [2]. In order to know if enhancement is suitable for L2 perceptual training, more research has to be done on enhancement and its effects on generalization, memory, and production.

ACKNOWLEDGMENTS

This work was funded by an EPSRC project grant (GR/L25639).

NOTES

1. The lack of significant effects for 'condition' in discrimination tasks could be related to a ceiling effect due to a learning effect (i.e., all subjects performed discrimination after identification tasks), and the use of extra-cues such as intonation, since the similar items of the AXB task were identical copies of the same token. A second experiment is being prepared where these two problems are solved.

REFERENCES

[1] Best, C.T. 1996. A direct-realist view of cross-language speech perception. in W. Strange (Ed.) *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues* (pp. 229-273). Timonium, MD: York Press.

[2] Bradlow, A., Pisoni, D., Yamada, R., Tohkura, Y. 1997. Training Japanese listeners to identify /r/ and /l/: IV. Some effects of perceptual learniong on speech production. *JASA*, 101 (4), 2299-2310.

[3] Hazan, V. and Simpson, A. 1998. The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication*, 24, 211-226.

[4] Jamieson, D. and Morosan D. 1986. Training non-native speech contrasts in adults: Acquisition of the English /ð/-/θ/ contrast by francophones. *Perception & Psychophysics*, 40 (4), 205-215.

[5] Jusczyk, P. 1997. *The discovery of spoken language*. Cambridge, Massachussets: MIT Press.

[6] Lively, S., Pisoni, D., Yamada, R., Tohkura, Y., and Yamada, T. 1994. Training Japanese listeners to identify English /r/ and /l/. III. Long term retention of new phonetic categories. *JASA*, 96 (4), 2076-2087.

[7] Logan, S., Lively, E. and Pisoni, D. 1991.Training Japanese listeners to identify English /r/ and /l/: A first report. *JASA*, 89 (2), 874-886.