Perception of English minimal pairs in noise by Japanese listeners: does clear speech for L2 listeners help?

Shinichi Tokuma[1], *Won Tokuma* [2]

[1] Department of English, Faculty of Commerce, Chuo University, Japan
[2]Faculty of Social Innovation, Seijo University, Japan

tokuma@tamacc.chuo-u.ac.jp

Abstract

This study investigates the effect of different types of clear speech on cross-linguistic perception. In particular, it examines whether the words produced to help L2 interlocutors are perceived more easily by them than the words produced for L1 interlocutors to counteract speech babble noise. English keywords in these conditions are extracted from the LUCID corpus, and they are presented to Japanese listeners for the identification of the words in a minimal pair, in babble noise or without noise. The results demonstrate that when presented in noise, the words produced to help L2 interlocutors are generally perceived more easily by them. This suggests that speech production varies according to listeners' needs in a specific communicative situation. The strong effect of the acoustic characteristics of Japanese /b/ on the perception by Japanese L1 listeners is also discovered. Further acoustic and perceptual research is required to investigate whether a universal or specific cue-enhancement is involved in the clear speech for L2 listeners.

Index Terms: L2 speech perception, clear speech, perception in noise

1. Introduction

When speakers realise that their interlocutors have difficulty understanding them, they adopt a speaking style called clear speech in order to enhance intelligibility for the presumed benefit of their interlocutors. Studies have shown that clear speech enhances intelligibility for both normal hearing and hearing-impaired listeners. (e.g. [1]) The acoustic realisation of this clear speech is characterised in many ways: segmentally by a more extended vowel space, larger contrast in VOT and duration, and prosodically by decreased speech rate, an increase in the number/duration of pauses and an increase in pitch range and mean F0 (see the papers reviewed in [2]). This is in agreement with Lindblom's H&H model [3], which proposes that speakers 'hyper-articulate' (i.e. increase the articulatory effort to produce clearer speech) to secure successful communication.

Clear speech is also employed when communicating with L2 interlocutors, especially in adverse listening conditions such as in noise. The speech perception of L2 listeners is not as robust as that of L1 listeners and they are more vulnerable to noise and reverberation [4][5][6]. Hence, L1 speakers modify their phonetic output to clarify their speech when they realise that their oral communication with their L2 interlocutors provides a challenge for them.

However, clear speech is not a unitary speaking style. Hazan and her colleagues [7][8] demonstrate that words produced to counteract speech babble noise are more easily perceived when presented in babble than words produced to counteract a vocoder (i.e. speech heard via a vocoder), which suggests that different types of hyper-articulation are rendered according to the types of adverse speech conditions. This raises the following question: does the L2 perception score differ according to the types of clear speech? Specifically, are the words produced to help L2 interlocutors perceived more easily by them than the words produced for L1 interlocutors to counteract other adverse conditions such as speech babble noise?

Although the effect of clear speech on L2 perception has been extensively studied, (e.g. [9]), little has been found on the effect of the types of clear speech on cross-linguistic speech perception. The aforementioned [7][8] examined L1 perception of three types of clear speech but all in L1 conditions. This study investigates whether the results of [7][8] hold for cross-linguistic perception.

To serve this purpose, speech materials intended to combat the L2 condition and babble noise (of the L1 condition) are used as the stimuli, and they are taken from the LUCID corpus [10]. The advantage of using the LUCID corpus is that it rendered a 'diapix' task, where two speakers talked to each other to find a number of differences between two versions of the same cartoon picture that contained twelve differences. Each speaker was seated in a separate sound-proof room and could not see the other party or her picture, and the only way to 'spot the difference' was to communicate via headsets with a microphone. This set-up ensured natural, spontaneous speech in different clear-speech conditions. Another advantage is that, by following the procedure of [7][8], which also utilised multi-talker babble materials from the LUCID corpus, the results of this study can be compared with their L1 data.

2. Experiment

2.1. Materials

In the LUCID corpus, the speakers A and B working in a pair completed the diapix task in three adverse speech conditions: when the speaker A's voice was passed through a three-channel noise-excited vocoder (Vocoder); mixed with multi-talker babble noise (Babble); and when the speaker B was a low-proficiency non-native English speaker (L2). The speakers were all native speakers of Southern British English, except in the L2 condition, where the speaker B's were native speakers of Chinese (Mandarin/Taiwanese) or Korean with a low English command. In this study, materials in the L2 condition were selected, as well as those in the Babble condition, whose L1 perception data is presented in [7][8]. The materials were taken from the speech of the speaker A.

The diapix tasks drawn from LUCID were designed to elicit a set of minimal or near-minimal pairs with sounds /b/-/p/ and /s/-/ʃ/. Although eight minimal or near-minimal pairs were chosen as stimuli in [7][8], this study utilised only two pairs: 'beach-peach' and 'sell-shell'. This was because all of the four words in the pairs are established as loan words in

Japanese, while the other pairs contain words that may not be familiar to some Japanese L2 learners, and this would affect their overall scores. In LUCID, the keywords were distributed over 12 diapix picture pairs and both keywords in a pair did not always appear in the same picture.

Due to the spontaneous nature of the diapix task, some speakers in the LUCID corpus did not utter all of the four words concerned: beach, peach, sell and shell. On other occasions, some of the words were whispered, produced in an extremely low intensity, or uttered as a part of a noun compound (e.g. seashell / shellfish). These speakers were excluded, and tokens from three speakers (one male, two females) were selected for stimuli. Since the speakers in the LUCID corpus had completed either the tasks of the L2 condition or the Babble condition [10], the words of the minimal pairs came from the identical speaker for each L2 / Babble condition, but not across the two conditions. Each token for the individual minimal-pair word was extracted from the corpus, and in a few cases where there was more than one token that was suitable for stimuli, one of them was selected randomly. All tokens were normalised to a fixed intensity level, using the Speech Filing System (ver. 4.8, obtainable from www.phon.ucl.ac.uk/resource/sfs/) and then further two conditions were made: one was to present the stimuli without babble noise (Clear) while the other was to present them with babble noise (Noise). In the Noise condition, 100-talker babble noise was added [11] at a signal-to-noise ratio of 0dB, again with the Speech Filing System. This process produced 48 stimulus tokens: 4 token types (beach / peach / sell / shell) x 3 speakers x 2 stimulus conditions (L2 / Babble) x 2 presentation conditions (Clear / Noise). See the diagram below for clarification.

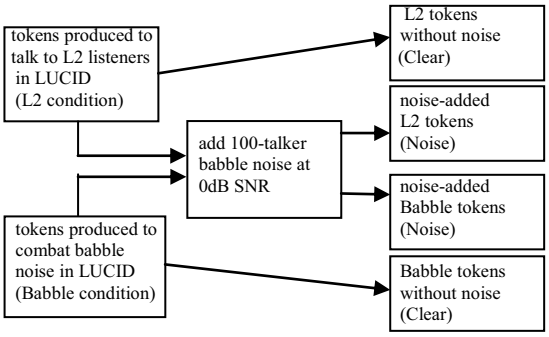

Figure 1: Stimulus manipulation.

2.2. Participants

Two groups of first-year Japanese undergraduate students at Chuo University in Tokyo, aged between 18 and 20, participated in the experiment. Each group belonged to a different English class, with 24 and 23 members, and the participants were tested on separate dates in the same Language Laboratory room (see details below). None of them had lived in an English-speaking country or reported on hearing / language impairment. They were all right-handed monolinguals. They had studied English as a foreign language for at least six years at school, and their English abilities were judged to be at the pre-intermediate level by their instructor. They were not paid for their participation.

2.3. Procedure

In [7][8], the participants were required to identify the initial consonant of the word presented. However, the task of this study was L2 perception, and to reduce their burden, they were instructed to identify the whole word.

The participants were tested in a quiet Language Laboratory room at Chuo University, in two separate groups as mentioned above, with the stimuli played through covered-ear headphones at a level adjusted by them. None of them reported that their attention had been diverted by noise, or by the presence of other participants. One group (24 members) listened to the stimuli, with or without noise, in the L2 condition, while the other group (23) listened to those in the Babble condition.

The participants were seated individually in front of laptop PCs, and their task was to listen to the stimulus token, with or without babble noise, through headphones and, using a mouse, click on the correct word in the minimal pair that appeared on the screen, as quickly and accurately as possible. The locations of the words on the screen, i.e. whether a word appeared on the right or left, were also randomised, and each stimulus token was played in a random order to the participants four times. In total, there were 96 presentations per participant: 4 token types (beach / peach / sell / shell) x 3 LUCID speakers (2 females and 1 male) x 2 presentation conditions (Clear / Noise) x 4 repetitions. The whole experiment process was controlled by Praat (ver. 5.3.34), utilising Experiment MFC objects.

The experiment was preceded by a task demonstration by one of the authors designed to make the participants familiar with the experimental setting and the nature of the stimuli. Care was taken to ensure that the participants knew that they must not click a word on a screen before the whole word was played, through instructions before and after the demonstration.

2.4. Results

In the analysis, the mean percentages of correct responses, as well as the mean reaction time (RT) for correct responses, were calculated for each token type, stimulus condition and presentation condition. During this process, seven participants, four in the L2 condition and three in the Babble condition, were excluded as a precaution, because RT in a negative value was found in their results. This reduced the total number of the participants to 40: 20 for each stimulus condition.

Figures 2 and 3 show the mean percentages of correct responses for each token type and presentation condition, i.e. whether babble noise was added (Noise) or not (Clear). Figure 2 is for the L2 condition (i.e. the interlocutor was an L2 speaker in the LUCID corpus) and Figure 3 for the Babble condition (i.e. the speaker's voice was mixed with babble noise in LUCID).

Overall, Figures 2 and 3 display a high rate of correct responses when the token was presented in the Clear presentation condition. However, when the babble noise was added (Noise condition), the percentages of correct responses decrease for all tokens in the Babble condition although the decrease rate is smaller or nearly zero in the L2 condition. This shows that for Japanese L2 listeners, perception of speech uttered in the L2 condition is generally more robust to noise than that uttered to overcome the babble noise. The exception to this is the initial /b/: in noise, it has a low correct score in both Babble and L2 conditions (68% and 66% respectively). In fact, the correct identification rate of /b/ in noise is lowest

across all stimulus and presentation condition, and this reflects the perceptual difficulty of English /b/ by Japanese L2 listeners.

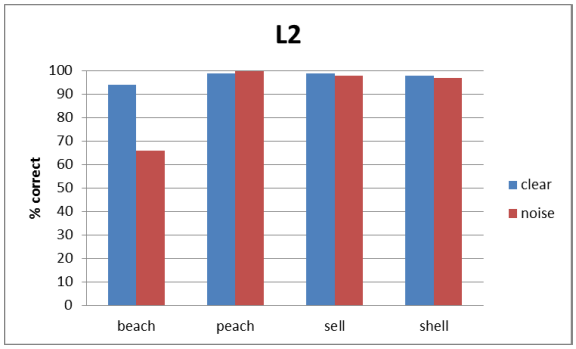

Figure 2: Mean identification scores for L2 condition.

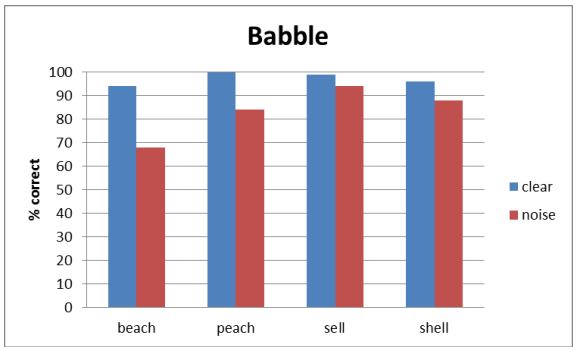

Figure 3: Mean identification scores for Babble condition.

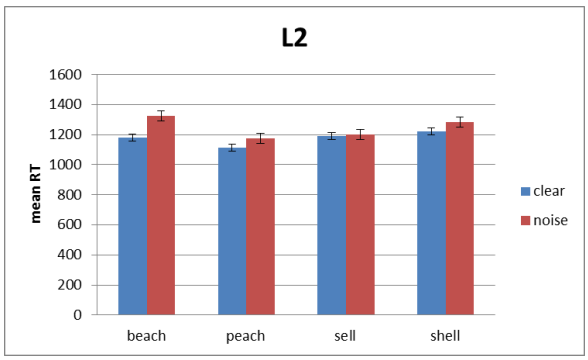

Figure 4: Mean RT for L2 condition.

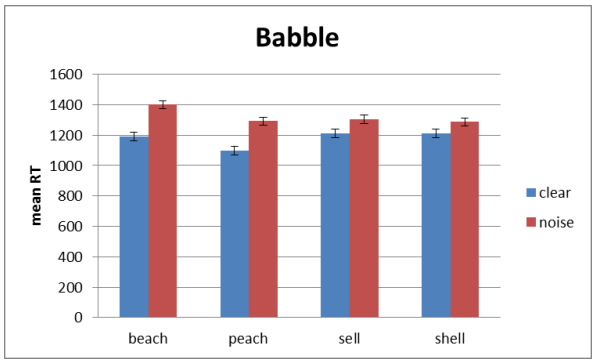

Figure 5: Mean RT for Babble condition.

In [7], the rate of correct responses for the tokens of the Babble condition which were presented in the Noise condition was 89% for /b/-/p/ and 95% for /s/-/ʃ/. In this study, if the average figure is calculated, it was 76% for /b/-/p/ and 91% for /s/-/ʃ/. This endorses the results of previous studies, which claim that L2 listeners are more susceptible to noise [4][5][6].

Figures 4 and 5 show the mean RT for each correctly-answered token type and presentation condition. Figure 4 is for the L2 condition and Figure 5 for the Babble condition. The error bars show 95% confidence intervals. The unit in the vertical axis of the Figures is milliseconds. The values of RT are generally much larger than those found in [7] due to the fact that it was an L2 perception task, as well as the difference in the procedure: participants were told to identify the initial consonant in [7], while their task was to identify the whole word in this experiment.

In Figures 4 and 5, the RTs for the Noise condition are slower than those for the Clear condition, endorsing the difficulty of L2 perception in noise. Also the mean RT differences between the Noise and Clear conditions are obtained, by calculating [mean value for Noise conditions]-[mean value for Clear conditions], for each L2 and Babble stimulus condition, and they are displayed in Table 1.

	beach	*peach*	*sell*	*shell*
L2	*144.5*	*65.1*	*9.7*	*60.1*
Babble	*208.3*	*195.95*	*93.9*	*75.7*

Table 1: Mean RT difference between Noise and Clear conditions, in ms.

Table 1 shows that the difference in RTs between Noise and Clear conditions is smaller for L2 stimulus condition, which suggests that perceiving clear speech for L2 listeners is more effective. The slowest RTs for /b/ in the Noise condition also reflect the perceptual difficulty of /b/ among Japanese L2 listeners mentioned above.

For the statistical analyses of the data, Model-Selection Loglinear model analysis was made on the number of correct responses, and the examined factors were the stimulus condition (L2, Babble), the presentation condition (Clear, Noise), and the token type (beach, peach, sell, shell). The final model produced is an interaction between the presentation condition and the token type ($p<.001$). Repeated measures ANOVA that was carried out on the RT showed a significant difference in the token type ($F(3,12)=4.78$, $p<.002$), the stimulus condition ($F(1,12)=3.07$, $p<.027$), and the presentation condition ($F(1,12)=24.15$, $p<.001$). The result of these analyses generally supports the observations above.

3. Discussion

The results of the experiment show that clear speech produced to help L2 interlocutors is generally perceived more easily by them than that produced for L1 interlocutors in speech babble noise. This is congruent with the results of [7][8], where words produced to counteract speech babble noise are more easily perceived when presented in babble than those produced to counteract vocoder speech. The results also suggest that different types of hyper-articulation are rendered according to the types of adverse speech conditions, as also claimed by [7][8]. This is in line with Lindblom's H&H model [3], which proposes that speech production varies according to listeners' needs in a particular communicative situation.

These observations raise the question of why speech in the L2 condition is perceived more easily in most cases. To

attempt to answer this, acoustic-phonetic analyses of the stimulus words that were used in the experiment were conducted. Note that the words were excised from spontaneous speech of the LUCID corpus, and they vary greatly in terms of their position in the sentence, context and prosodic contour.

The mean duration values were calculated for each token type and stimulus condition (L2 / Babble), and they are displayed in Table 2 below.

	beach	*peach*	*sell*	*shell*
L2	*379.0*	*285.7*	*439.6*	*490.5*
Babble	*416.3*	*313.1*	*363.1*	*417.8*

Table 2: Mean duration in ms.

Table 2 shows that the mean duration of 'sell' and 'shell' is longer in the L2 condition than that in the Babble condition, suggesting that slower speech rate, realised as a longer duration, could help L2 listeners. However, the mean duration of 'beach' and 'peach' is shorter in the L2 condition. This means that the better correct perception rate can be attributed to the other acoustic cue. To examine this, VOT measurements of the initial plosives, as well as of the mean F0 range, were made and their mean values are shown in Tables 3 and 4. The F0 range is considered to be one of the properties of the voicing distinction (for example, see [12]), and it was calculated in semitones relative to 1 Hz by Praat.

	beach	*peach*	*difference*
L2	*-2.6*	*39.3*	*41.9*
Babble	*-4.0*	*52.6*	*54.6*

Table 3: Mean VOT in ms.

	beach	*peach*	*difference*
L2	*2.5*	*4.7*	*2.2*
Babble	*3.0*	*4.1*	*1.1*

Table 4: Mean frequency range in semitones.

In Table 3, the mean VOT difference between /b/ and /p/ is 54.6ms for the Babble condition but it is 41.9ms for the L2 condition, and therefore the enhancement of VOT difference was not made between two conditions. In contrast, the difference in mean F0 range in Table 4 is 2.2 semitones for the L2 condition while it is 1.1 semitones for the Babble condition, suggesting that the increase in the F0 range could contribute the perceptual enhancement of the /b/ and /d/ in the L2 condition. This implies that a different phonetic cue enhancement could have been made according to the type of segments. However, this study has not explored the possible involvement of other less prominent cues and 'global adjustments' [9], which increase the overall salience of the signal, and further acoustic and perceptual experiment is required to look into this matter more specifically.

There is one further issue to be discussed here. Figures 2 and 3 show that the mean correct percentage of /b/ is lower than that of /p/ in all cases, which indicates that Japanese listeners have more difficulty in perceiving English /b/ than /p/. This is not due to a lack of phonological /p/-/b/ contrast in Japanese; an example of the distinction is /paɴ/ 'bread' – /baɴ/ 'order/turn'. Rather, it can be attributed to the difference in the acoustic realisation of /b/ in both languages. In Japanese, a word initial /p/ is weakly aspirated [13] and it should be easy for Japanese listeners to identify its English counterpart, which is strongly aspirated. On the other hand, Japanese /b/ has no aspiration, and it occasionally becomes a fricative or an approximant when flanked by vowels [14], which rarely happens in English. Acoustically, in the word-initial position, Japanese /b/ is frequently realised often with long pre-voicing. For example, [12] states that the average VOT for /b/ in the initial position is -89ms, while that for /p/ is 41ms. In contrast, long pre-voicing is very rare in the word-initial position in English. Hence, it may be inferred that Japanese listeners may find it more difficult to identify English /b/, which lacks strong pre-voicing. This can be supported by a slightly higher correct identification score of /b/ in Babble, 68%, (where average VOT is -4.0ms) than in L2, 66% (where it is -2.6ms). Needless to say, this is a working hypothesis and requires further verification.

4. Conclusion

The results demonstrate that when presented in noise, clear speech tokens produced to help L2 interlocutors are perceived more easily by them than those for L1 interlocutors to counteract noise. This suggests that speech production varies according to listeners' needs in a specific communicative situation. Further acoustic and perceptual research is required to investigate whether a universal or specific cue-enhancement is involved in the clear speech for L2 listeners. The strong effect of the acoustic characteristics of Japanese /b/ was also discovered on the perception by Japanese L1 listeners.

5. Acknowledgements

Thanks are due to Valerie Hazan of University College London, who provided the access to the LUCID corpus. The help on Speech Filing System by Mark Huckvale of University College London is also highly appreciated, as well as the comments from Fumio Hirasaka of Kanto Gakuin University. This research is partially funded by Chuo University Grant for Special Research.

6. References

[1] Liu, S., Del Rio, E., Bradlow, A.R. and Zeng, F.-G., "Clear speech perception in acoustic and electrical hearing." Journal of the Acoustical Society of America, 116: 2374-2383, 2004.

[2] Granlund, S., Hazan, V. and Baker R., "An acoustic-phonetic comparison of the clear speaking styles of Finnish-English late bilinguals." Journal of Phonetics, 40: 509-520, 2012.

[3] Lindblom, B., "Explaining phonetic variation: a sketch of H&H theory," in *Speech production and speech modelling*, W.J. Hardcastle and A. Marchal, eds. Amsterdam: Kluwer Academic, 1990, pp. 403-439.

[4] Florentine, M., "Non-native Listeners' Perception of American-English in Noise," Proceedings of Inter-Noise '85: 1021-1024, 1985.

[5] Takata, Y. and Nabelek, A.K., "English consonant recognition in noise and in reverberation by Japanese and American listeners." Journal of the Acoustical Society of America, 88: 663-666, 1990.

[6] Garcia-Lecumberri, M.L. and Cooke, M., "Effect of masker type on native and non-native consonant perception in noise." Journal of the Acoustical society of America, 119: 2445-2454, 2006.

[7] Grynpas, J., Baker, R. and Hazan, V., "Clear speech strategies and speech perception in adverse listening conditions." Proceedings of XVIIth ICPhS, Hong Kong, 2011: 779-782, 2011.

[8] Hazan, V., Grynpas, J. and Baker, R., "Is clear speech tailored to counter the effect of specific adverse listening conditions?" Journal of the Acoustical society of America, On-Line Express Letter. 132: EL371-EL377, 2012.

[9] Bradlow, A.R. and Bent, T., "The clear speech effect for non-native listeners." Journal of the Acoustical society of America, 112: 272-284, 2002.

[10] Baker, R. and Hazan, V., "DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs." Behaviour Research Methods, 43: 761-770, 2011.
[11] Varga, A. and Steeneken, H.J.M., "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems." Speech Communication, 12: 247-251, 1991.
[12] Shimizu, K., *A cross-language study of voicing contrasts of stop consonants in Asian languages*. Tokyo: Seibido, 1996.
[13] Riney, T.J., Takagi, N., Ota, K. and Uchida, Y. "The intermediate degree of VOT in Japanese initial voiceless stops." Journal of Phonetics, 35: 439-443, 2007.
[14] Vance, T.J., *The Sounds of Japanese*. Cambridge: Cambridge University Press, 2008.