JAPANESE LISTENERS' IDENTIFICATION OF ENGLISH VOICELESS FRICATIVES IN REVERBERANT LISTENING ENVIRONMENTS

Hinako Masuda

Faculty of Science and Engineering, Waseda University, Japan
h-masuda@aoni.waseda.jp

ABSTRACT

This research investigated the identification ability of English voiceless fricatives by Japanese listeners with high and low English proficiency. A perceptual experiment was carried out in five listening environments: RT (Reverberation Time) = 0.78 s, 1.12 s, 1.43 s, RT = 0.78 s + background noise at SNR (Signal-to-Noise ratio) = 10 dB, and quiet. Correct identification rates were calculated and influence of the Japanese listeners' English proficiency was considered by means of TOEIC® scores. In addition, confusion matrices were created to investigate the misperception patterns. The results were then compared with those of native English listeners'. Results showed that there was a significant difference between English and Japanese listeners, but no significant difference between English listeners and Japanese listeners with higher English proficiency. However, detailed analyses of misperception patterns revealed differences between the two groups as well as similarities between Japanese with higher and lower English proficiency.

Keywords: L2 speech perception, consonants, proficiency, background noise, reverberation

1. INTRODUCTION

Speech perception in real-life environments is almost always accompanied by background noise and reverberation. Despite this fact, non-native listeners are often trained to listen to foreign sounds in a quiet, laboratory environment which does not reflect real-life environments at all. Based on the premise that the aim of L2 speech perception training is to gain the ability to perceive foreign sounds in real-life listening environments, i.e., in background noise and/or reverberation, it is important for us to know the difficulties that non-native listeners face in such situations.

Speech perception in noisy and reverberant environments is challenging for all listeners, both native and non-native. Needless to say, the challenge is bigger for non-native listeners. Past research [4, 8, 9] has demonstrated that even non-native listeners with high fluency in the target language, such as early bilinguals, fell short of native listeners when sounds were presented in background noise or reverberation. However it has not been specifically clarified what high proficiency listeners are having difficulty with. This paper is a report on a case study focusing on English voiceless consonants and is a part of a larger project which in whole aims to understand the mechanism of the perception of English sounds in background noise and reverberation by Japanese listeners, and to make use of the data for developing perceptual training materials. The project is particularly interested in how one's foreign language proficiency affects perception, and in clarifying the difficulty of listeners with high proficiency. This is explored by looking at the correct identification rates as well as misperception patterns.

According to the International Phonetic Association [5], there are 24 consonants in North American English. The breakdown of the 24 consonants are: six plosives /p b t d k g/, two affricates /ʧ ʤ/, three nasals /m n ŋ/, nine fricatives /f v θ ð s z ʃ ʒ h/, three approximants /ɹ j w/, and one lateral approximant /l/. The Japanese phonetic system, on the other hand, has a total of 16 consonants: six plosives /p b t d k g/, one affricate /tˢ/ (or /tɕ/, which comes before /i/ and /u/), three nasals /m n ɴ/, one flap /ɽ/, three fricatives /s z h/ (phonetically, seven fricatives [ɸ s z ɕ ʑ ç h] [11]), and two approximants /j w/. It is worth noting that some of the sounds that seem to be the same (e.g. /t d n/) in the two systems actually have different places of articulation, i.e. differences in the acoustic features. For example, voiceless anterior stop consonants /t d n/ and voiceless sibilants /s ʃ/ appear similar in English and Japanese when in fact they are different in terms of place of articulation. /t d n/ are alveolar consonants in English; on the contrary, they are dentals in Japanese. Similarly, English /s/ is an apico-alveolar sound whereas Japanese /s/ is a lamino-dental sound. Also, /ʃ/ is a post-alveolar fricative sound in English, and alveolo-palatal sound in Japanese, expressed as [ɕ] [10]. While it is natural that some sounds overlap in two language systems, considering that sounds are created in a human vocal tract after all, the interest here is how such sounds are perceived by native and non-native, naïve listeners.

In the Perceptual Assimilation Model (PAM), Best [1] claims that unfamiliar sounds are assimilated to a native category, assimilated as uncategorizable speech sound, or not assimilated to speech sound at all and categorized as non-speech sound. The Japanese listeners' perception of English consonants is assumed to fit into the first and second categories. In view of this model and the findings from the author's previous study, we hypothesize that 1) Japanese listeners' performance is always lower than the English listeners regardless of listening environments, 2) when Japanese listeners are further divided into sub-groups depending on their English proficiency, the higher proficiency group will perform similar to the English listeners in terms of correct identification rates, 3) adverse listening conditions will depict the influence of the Japanese listeners' first language, i.e. the misperception patterns of Japanese listeners with higher English proficiency will resemble that of the lower proficiency. Due to space limitations, this paper will remain at reporting the misperception patterns of the listeners and not the detailed analyses of underlying acoustic evidence such as acoustic cues non-native listeners use to identify non-native sounds. This issue is to be addressed in future research.

2. PERCEPTUAL EXPERIMENT

A perceptual experiment was carried out to assess the Japanese listeners' ability to identify English voiceless fricatives in reverberant listening environments.

2.1. Participants

Twenty-two Japanese listeners participated in the perceptual experiment. All were recruited at a university in Japan, and received a small compensation as a reward for participating in the experiment. As a baseline for comparison, twenty American English listeners also participated in the experiment. All were recruited at a university in the United States, and their participation was voluntary. None of the participants reported any hearing problems.

2.2. Participants

Participants were presented with twenty-three English consonants /b ʧ d f g h ʤ ʒ k l m n p ɹ s ʃ t θ ð v w j z/ (excluding /ŋ/ from the 24 North American English consonants) embedded in the context "You are about to hear a__a". Among the 23 consonants, the present paper reports the result for five voiceless fricatives /f h s ʃ θ/. The stimuli were produced by a female Japanese-English bilingual speaker, and were recorded in a sound-attenuated chamber using a digital sound recorder (Marantz PMD 660) and a microphone (SONY ECM-23F5) at a sampling frequency of 48 kHz. The sound files were later downsampled to 16 kHz.

All participants listened to the stimuli in the order of 1) reverberant environments of RT = 0.78 s (D50 value 67.5%), 1.12 s (D50 value 47.7%), and 1.43 s (D50 value 32.2%) in randomized order), 2) noisy and reverberant condition (SNR = 10 dB added to reverberation RT = 0.78 s), and 3) the quiet condition. Impulse responses for reverberant environments were taken from a reverberation corpus which were recorded at the NHK hall in Tokyo, Japan (RT = 0.78 s and 1.12 s) and Kamakura Museum of Art in the Tokyo metropolitan region (RT = 1.43 s). Multispeaker babble noise was selected as background noise [12] among others such as white noise, because 1) of the similarity of its long spectra and temporal variation to that of human speech, and 2) Lecumberri & Cooke [7] showed that multispeaker babble noise depict greatest difference between native and non-native speech perception in background noise.

2.3. Procedure

A laptop computer was used to present the stimuli and to record the listeners' responses. All experimental procedure was conducted using Praat [2]. Stimuli were presented to the Japanese listeners through an USB audio amplifier (ONKYO MA-500U or Roland Duo-Capture EX UA-22) and headphones (Sennheiser HDA200 or SONY MDR CD900ST). The laptop computer and audio amplifier were digitally connected via USB interface. English listeners were presented with the stimuli through Sennheiser HD 280 Pro headphones connected directly from Mac computers.

Before the main experiment, both native and non-native participants went through a practice session which consisted of 23 practice trials. The main experiment consisted of 575 trials (23 consonants x 5 repetitions x 5 listening environments) per participant. Participants listened to each stimulus and were asked to choose one of the consonants out of the 23 choices that was most similar to what they heard, for example the correct choice for /aba/ would be "B as in Be", "CH as in Chin" for /ʧa/, etc (reference: [3]). Participants were presented with the stimuli only once, and trials automatically proceeded to the next trial after pressing the answer button on the computer screen.

Reaction time was not recorded. The experiment took approximately 60 minutes to complete, including the explanation of the experiment and filling out the questionnaire. For this particular paper, a total of 5250 trials (42 participants x 5 consonants x 5 repetitions x 5 listening environments) underwent analyses.

3. RESULTS & DISCUSSION

3.1. Mean identification rates

The mean identification rates of English and Japanese listeners are shown in Tables 1 and 2, respectively. Identification rates below 60% are shaded in gray. Taking away the effect of English proficiency of the Japanese listeners, two-factorial Analysis of Variance between subjects found significant differences in the identification rates of the English and Japanese listeners [$F(1, 200) = 10.1$, $p < 0.01$] and listening environments [$F(4, 200) = 56.2$, $p < 0.001$], but no interaction between the two [$F(4, 200) = 1.6$, $p = 0.15$]. *Post-hoc* comparisons using Tukey-Kramer's test showed significant differences ($p < 0.01$) in all but Quiet vs RT = 0.78s, RT = 0.78s vs RT = 1.12s and RT = 1.12s vs RT = 1.43s environments.

Overall, English listeners achieved over 65% in all except the most adverse listening environment (RT = 0.78s + SNR = 10 dB), except for /θ/ in which identification rate dropped to under 60% from RT = 1.12s. For the Japanese listeners, identification rates started to drop for /f s θ/ from RT = 1.12s. They maintained their native-like high performance of /h/ and /ʃ/ until the most adverse environment. /θ/ is often regarded as a difficult consonant to perceive for Japanese as it does not exist in Japanese, however the difficulty was also observed in English listeners.

Table 1: Mean identification rates of English listeners (%).

	f	h	s	ʃ	θ
Quiet	95.8	97.5	89.2	90.8	75.0
RT=0.78s	90.0	99.2	65.8	80.8	68.3
RT=1.12s	77.5	92.5	69.2	86.7	59.2
RT=1.43s	76.7	89.2	74.2	93.3	58.3
RT=0.78s+SNR =10dB	30.8	58.3	43.3	60.0	55.8

Table 2: Mean identification rates of all Japanese listeners (%).

	f	h	s	ʃ	θ
Quiet	99.1	100	82.7	97.3	58.2
RT=0.78s	81.8	100	60.0	90.0	64.5
RT=1.12s	60.0	96.4	39.1	86.4	54.5
RT=1.43s	64.5	93.6	44.5	96.4	58.2
RT=0.78s+SNR =10dB	24.5	41.8	21.8	51.8	20.9

3.2. English proficiency of the Japanese listeners

Japanese listeners are further divided into two groups with respect to TOEIC® (Test of English for International Communication) scores. The higher Japanese listeners (HJ, $N = 10$; Table3) are those who achieved higher than 720, and the lower Japanese listeners (LJ, $N = 12$; Table 4) are those who achieved lower than 600. Two-factorial Analysis of Variance between subjects showed significant differences in both listener groups [$F = (2, 195) = 19.1$, $p < 0.001$] and listening environments [$F = (4, 195)$, $p < 0.001$] but no interaction between the two [$F = (8, 195)$, $p = 0.22$]. *Post-hoc* comparisons using Tukey-Kramer test showed significant differences in the identification rates between English listeners and LJ, as well as HJ and LJ, but not between English listeners and HJ. That is, in terms of identification rates, Japanese listeners with higher TOEIC® scores performed similarly to English listeners whereas those with lower TOEIC® scores achieved significantly lower compared to both English listeners and HJ.

Table 3: Mean identification rates of HJ (%).

	f	h	s	ʃ	θ
Quiet	98.0	100	84.0	98.0	78.0
RT=0.78s	86.0	100	74.0	100	92.0
RT=1.12s	70.0	96.0	52.0	98.0	76.0
RT=1.43s	74.0	98.0	60.0	98.0	80.0
RT=0.78s+SNR =10dB	28.0	30.0	22.0	68.0	40.0

Table 4: Mean identification rates of LJ (%).

	f	h	s	ʃ	θ
Quiet	100	100	81.7	96.7	41.7
RT=0.78s	78.3	100	48.3	81.7	41.7
RT=1.12s	51.7	96.7	28.3	76.7	36.7
RT=1.43s	56.7	90.0	31.7	95.0	40.0
RT=0.78s+SNR =10dB	21.7	51.7	21.7	38.3	5.0

3.3. Confusion matrices

Japanese and English listeners' misperception patterns showed both similarities and differences. Both English and Japanese listeners suffered in perceiving /θ/ accurately. The closest research to the present report is one by Lambacher *et al.* [6] in which Japanese and English listeners were presented with five English voiceless fricatives /f s ʃ θ h/ in CV, CV and VCV contexts in a quiet listening environment. The English listeners in their study achieved close to a perfect score for the identification of /θ/, whereas the Japanese listeners achieved 55.2%. Although the English listeners' data in the present paper falls back by approximately 20% compared to those in Lambacher *et al.*'s study, the results of the Japanese listeners were virtually identical. Thus the results of this paper confirmed the disadvantage of the Japanese listeners with relatively low English proficiency in the identification of /θ/.

The reverberant listening conditions of the present experiment gained new understanding that the English listeners' difficulty of accurately perceiving /θ/ increases until it is virtually identical to the Japanese listeners' results in the RT = 1.12s and 1.43s listening environments. The misperception patterns, however, are quite different between the two listener groups. While English listeners mostly misperceived /θ/ as /ð/, misperception as /f/ appeared as listening environments became more adverse (namely RT = 1.43s and RT = 0.78s + SNR = 10dB). Japanese listeners, on the other hand, rarely misperceived /θ/ as /ð/ but mostly /f/. An interesting point is that the HJ group's performance surpassed that of the English listeners; HJ did not have difficulty in accurately perceiving /θ/ except for the most adverse listening environment.

Japanese listeners also had a disadvantage on the identification of /s/. This time, HJ had difficulty in accurately identifying this consonant just as the LJ group, although not as severe. English listeners did not have difficulty until the listening environment was most adverse, in which case they misperceived it as /z/, a voiced counterpart of the target sound. The HJ and LJ groups also had the tendency to misperceive /s/ as /z/ but also as /θ/ and /ð/. The identification of /f/ was similar in English listeners and the HJ group in terms of identification rates, but the misperception patterns differed. English listeners misperceived /f/ as /p/ most of the time, whereas the HJ group mostly as /θ/ and some /p/, and LJ as /h/, /p/, and /t/. As for the consonants /h/ and /ʃ/, all three listener groups identified them well up to the most adverse listening environment. All groups misperceived /ʃ/ as /ʧ/. English listeners misperceived /h/ as /p/, whereas both HJ and LJ groups misperceived /h/ as not only /p/ but also /f/ and /m/.

By taking a close look at the misperception patterns of the English and Japanese listeners and taking further consideration of the Japanese listeners' English proficiency, similarities and differences become highlighted. Such tendencies shed light in setting priorities in which consonants need to be trained in adverse listening environment to bridge the gap of non-native disadvantage and develop a perceptual training material for the Japanese learners of English.

4. CONCLUSION

The present paper reported the Japanese listeners' identification ability of English voiceless fricatives in reverberant listening environments, compared the results with that of the English listeners', and looked further into their misperception patterns. Results indicated that in terms of identification rates, Japanese listeners' performance was significantly lower than that of English listeners. However, when Japanese listeners were further divided into sub groups with respect to their TOEIC® scores, Japanese listeners with higher scores performed similarly to the English listeners. However, misperception patterns revealed the differences between English listeners and HJ, as well as similarities between HJ and LJ groups. These misperception patterns in various listening environments that mock real-life listening situations shed light on how perceptual training materials should be developed.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant numbers 24820043 (Grant-in-Aid for Research Activity Start-up) and 2658011 (Grant-in-Aid for Challenging Exploratory Research), and Waseda University Grants for Special Research Projects (Grant number 2014S-100). The author would like to thank Takayuki Arai and Shigeto Kawahara for their comments, and Hideki Tachibana, Kanako Ueno and Sakae Yokoyama for offering to use the impulse response data for simulating the reverberation listening environments. The author would also like to thank the research assistants at the Rutgers University phonetics laboratory for assisting in collecting the English listeners' data.

5. REFERENCES

[1] Best, C.T. 1995. A direct realist view of cross-language speech perception. In: *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, 171-204.

[2] Boersma, P., Weenink, D. 2014. *Praat: doing Phonetics by computer* [Computer program]. Version 5.3.66, retrieved 24 March 2014 from http://www.praat.org/

[3] Cutler, A., Weber, A., Smits, R., Cooper, N. 2004. Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America* 116 (6), 3668-3678.

[4] Florentine, M. 1985. Non-native listeneres' perception of American-English in noise. *Proceedings of Inter-noise* 85, 1021-1024.

[5] International Phonetic Association, *Handbook of the International Phonetic Alphabet –A guide to the use of the International Phonetic Alphabet*-, Cambridge University Press (1999).

[6] Lambacher, S., Martens, W., Nelson, B., Berman, J. 2001. Identification of English voiceless fricatives by Japanese listeners: Influence of vowel context on sensitivity and response bias. *Acoustical Science and Technology* 22 (5), 334-343.

[7] Lecumberri, M.L.G., Cooke, M. 2006. Effect of masker type on native and non-native consonant perception in noise. *Journal of the Acoustical Society of America* 119(4), 2445-2454.

[8] Mayo, L.H., Florentine, M., Buus, S. 1997. Age of second-language acquisition and perception of speech in noise. *Journal of Speech, Language, and Hearing Research*, 686–693.

[9] Rogers, C.L., Lister, J.J., Febo, D.M., Besing, J.M., Abrahams, H.B. 2006. *Applied Psycholinguistics* 27, 465–485.

[10] Toda, M., Honda, K. 2003. An MRI-based cross-linguistic study of sibilant fricatives. *Proceedings of the 6th International Seminar on Speech Production*, 1-4.

[11] Vance, T.J. 1987. *An Introduction to Japanese Phonology*. State University of New York Press, New York.

[12] Varga, A., Steeneken, H.J.M. 1993. Assessment for automatic recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication* 12 (3), 247-251.