ICSLP, 3:750-753, 2000

THE EFFECT OF REDUCED SPECTRAL INFORMATION ON JAPANESE CONSONANT PERCEPTION: COMPARISON BETWEEN L1 AND L2 LISTENERS

Masahiko Komatsu[1,2], Won Tokuma[3], Shinichi Tokuma[4], Takayuki Arai[1]

[1] Sophia University, Tokyo, Japan
http://www.splab.ee.sophia.ac.jp
[2] University of Alberta, Edmonton, AB, Canada
[3] Seijo University, Tokyo, Japan
[4] Sagami Women's University, Sagamihara, Kanagawa, Japan

ABSTRACT

We investigated how spectral information contributes to the perception of Japanese consonants, using re-synthesised samples that were created by (1) gradually reducing the order of LPC analysis in the residual excited LPC vocoder; and (2) gradually flattening the spectral peak in the frequency domain. The results of native Japanese speakers showed that the information in LPC residuals contributes significantly, if not sufficiently, to Japanese consonant perception, and that the minimum amount of spectral information is sufficient to achieve 90% identification score. It was also found that, although the perceptual error patterns were different, there were striking similarities between Japanese and non-Japanese listeners in their averaged perception scores. The phonological feature analysis of the perceptual results indicated that the residuals provide broad phonotactic information such as major class features.

1. INTRODUCTION

LPC analysis separates speech signals into residuals (representing source information) and coefficients (representing filter information). In linguists' term, the source information roughly corresponds to suprasegmental/prosodic information while the filter information roughly to segmental information.

Mori et al. [1] studied the role of prosodic information played in language identification, using LPC residuals as perceptual stimuli. In their experiment, the subjects could often detect some of the words in residuals, despite the lack of spectral information. Mori et al. assumed that the acoustic information which was present in the residuals (presence/absence of harmonic structure and noise component, temporal intensity change, and so on) may have served as perceptual cues for vowel/consonant distinction and manner of articulation.

In this paper, we examined:

1. whether LPC residuals are sufficient for listeners to recognise Japanese consonants;
2. if not, how much spectral information is necessary to improve the identification score.

It is also known that L2 learners' speech recognition is not as robust as that of L1 speakers. For example, Florentine [2] reported that the addition of background noise has more significant effects on speech perception of L2 learners than that of L1 speakers. Thus, it was thought to be prudent to test the above questions (1) and (2) on L2 learners.

Furthermore, the perceptual results were phonologically analysed to elucidate how the reduction of spectral information influences the perception of phonological features, particularly major class features, and to what extent the perception of these features are dependent upon source or filter information.

2. EXPERIMENT 1: EFFECTS OF THE NUMBER OF SPECTRAL PEAKS

2.1 Subjects

10 native speakers of Japanese (8 males and 2 females; 21-26 years old) and 3 native speakers of English learning Japanese as L2 (1 male and 2 females; 21 years old) participated in the experiment. The L2 learners had learned Japanese for 1.5-2.5 years since at the age of 18-20, and stayed in Japan for 1.4-8 months.

2.2 Stimuli

As original samples, 17 Japanese /C/+/a/ syllables（/ka/, /ga/, /sa/, /za/, /sha/, /ja/, /ta/, /da/, /cha/, /na/, /ha/, /ba/, /pa/, /ma/, /ya/, /ra/, /wa/） were selected from ATR Japanese database [3] and down-sampled to 16 kHz. From these original samples, 5 further stimulus sets were created by calculating their 22nd, 10th, 6th, 4th and 2nd order LPC coefficients (using Hamming window; frame: 512 points, 75% overlap) and applying these coefficients to its 22nd LPC residual. Their 22nd LPC residuals were also included in the stimulus sets. Thus 7 sets of stimuli were presented to the subjects: the original syllable set, five LPC re-synthesised sets, and the 22nd LPC residual set. The power of the residuals and re-synthesised syllables were adjusted at each frame to match their original samples.

2.3 Procedures

The 7 stimulus sets were presented to the subjects in the following order: the 22nd LPC residual set, five LPC re-synthesised sets (in the order of coefficients, i.e. 2/4/6/10/22), and the original syllable set. Each set had 51 stimulus presentations (17 syllable types x 3 repetitions) with a pause of 2 seconds between each stimulus. The presentation order was randomised. The subjects were asked to identify the /Ca/ syllables by selecting the right token written on an answer sheet.

The stimuli were presented through headphones, in the soundproof room of our laboratory at Sophia University.

2.4 Results and Analyses

The mean identification rates for each stimulus type and subject group (Japanese/English listeners) were calculated and are shown in Figure 1. In Figure 1, "E" represents residuals, "X" the original samples and "S*n*" re-synthesized samples with the *n*th order LPC coefficients (e.g., S2 corresponds to samples with 2nd order coefficients).

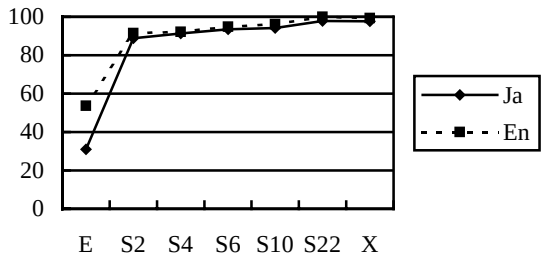


Figure 1: Percentage correct identification of consonants.

The Japanese subjects showed on average 39% identification rate for residuals, which is well above chance level. The identification rate rose to 89% for S2. The English subjects showed the similar tendency in mean identification rates: 54% for E and 92% for S2, although their error patterns were different from those of the Japanese subjects.

These results suggest that the information in the residuals contributes significantly to Japanese consonant perception, and that the addition of a single resonance peak was sufficient for Japanese listeners to achieve 89% identification rates. This was also the case for L2 listeners.

3. EXPERIMENT 2: EFFECTS OF SPECTRAL FLATTENING

3.1 Subjects

15 native speakers of Japanese (4 males and 11 females; 19-22 years old) and 4 native speakers of English learning Japanese as L2 (2 males and 2 females; 21-23 years old) participated in the experiment. The L2 learners had learned Japanese for 0.7-4 years since at the age of 16-21, and stayed in Japan for 4-10 months.

3.2 Stimuli

Using the same original samples as in Experiment 1, their 2nd order LPC coefficients and 22nd LPC residuals were obtained in the identical procedure, and in their LPC re-synthesis, the 2nd order LPC coefficients were manipulated by multiplying the magnitude of the original 2nd order LPC coefficients by 7 factors (1.00/0.95/0.90/0.80/0.60/0.40/0.00). This stimulus continuum was designed to replenish the gap between the syllables with the 2nd order LPC coefficients and the 22nd residuals, which showed a sharp rise in identification rates in Experiment 1. Reducing the factor flattens the spectral peak of the re-synthesised syllables: the factor 1.00 did not influence the LPC re-synthesis, while the factor 0.00 nullified the LPC coefficients and the residuals were obtained as the output of the re-synthesis. Thus 7 different stimulus sets were made according to 7 different factors. In this experiment, the re-synthesised samples of factor 0.00 had a spectral tilt of -6dB/octave.

3.3 Procedures

The 7 stimulus sets were presented to the subjects in the order of the factor size (i.e., from stimuli with 0.00 factor to those with 1.00 factor). As in Experiment 1, each set had 51 stimulus presentations (17 syllable types x 3 repetitions). The presentation order was randomised. In this experiment, the subjects were asked to identify the /Ca/ syllables which they listened to through headphones, by clicking with the mouse the right syllable appearing on a PC display. This experimental setup allows subjects to proceed at their own pace. This experiment was carried out in a CALL system room at Sophia University, which can accommodate several subjects at one time.

3.4 Results and Analyses

The mean consonant identification rates of Experiment 2 for each stimulus type and subject group were obtained and are shown in Figure 2, together with the error of one standard deviation (shown in dotted lines). In Figure 2, "S2_*nnn*" represents re-synthesized samples with the multiplied factor of *n.nn* (e.g., S2_040 corresponds to samples with the multiplied factor of 0.40).

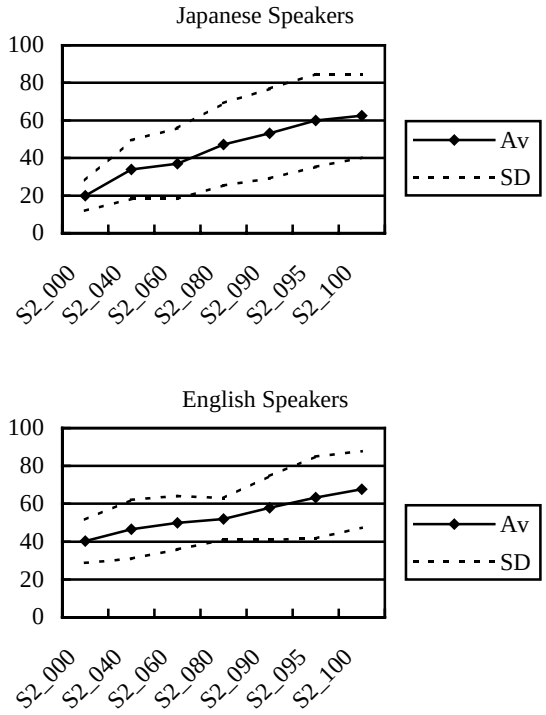


Figure 2: Percentage correct identification of consonants.

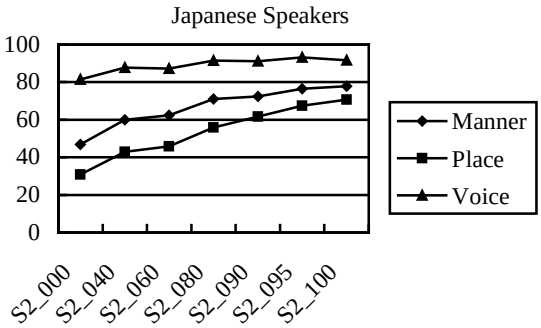
In Figure 2, the overall mean identification rates were lower than those observed in Experiment 1, and presumably this was due to the different listening environment. Figure 2 shows that the identification rates improved gradually from S2_000 to S2_100, but the large standard deviations in Figure 2 indicates that there was a large variance of identification scores among subjects. In fact some participants hardly showed improvement in identification rates from S2_000 to S2_100. As in Experiment 1, English speakers showed higher identification rates than Japanese speakers.

Consequently, the results are now analysed according to the phonological features of consonants, to study how the reduction of spectral information influences the perception of their features. Consonants are traditionally classified by the distinctive features: manner of articulation, place of articulation, and voice. We referred to Inozuka and Inozuka [4] for the definition of the manners and places of articulation. The identification score of each feature was obtained as follows: if the stimulus /ka/ was heard as /ga/, the manner feature [plosive] and place [velar] were regarded as being correct while the voice [-voice] as incorrect. Then the mean identification rates were calculated for each feature category (place/manner/voice), stimulus type and subject group, and they are shown in Figure 3.

In Figure 3, the rates of manner features are consistently higher than those of place features, which reflects the difference in human perceptual strategy between these features. Voice features show high identification rates throughout the stimulus types.

O'Shaughnessy [5] claims that in human speech perception process, each distinctive feature is perceived in a different way: the perception of manner of articulation mainly depends on gross energy distribution on spectrum and periodicity, while the perception of place depends on finer aspects of spectrum and complex interaction of several cues. This implies that the manner features are more robust to reduced spectral information of stimuli used in Experiment 2. O'Shaughnessy's claim also accounts for the phenomenon observed in Experiment 2 that the manner features were more accurately identified than the place features.

Figure 3 also indicates that again there were similarities between the Japanese and non-Japanese speakers in their perception scores. This suggests that native and non-native speakers of Japanese adopt the identical strategy for phonological feature perception.



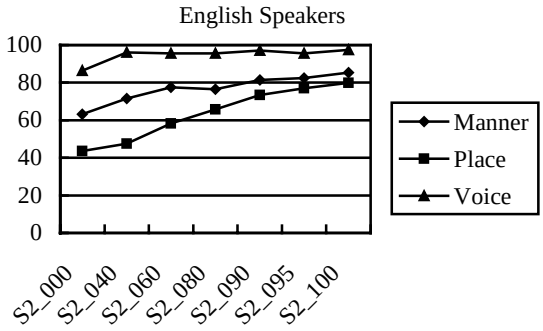


Figure 3: Percentage correct identification of features.

3.5 Phonological Analyses

The results were analysed with a major class feature model, which stipulates that consonants are classified as shown in Table 1 (See Kenstowicz [6]), each class being specified by binary features: [consonantal] and [sonorant]. In this study the feature [approximant] was added to distinguish nasals from liquids, although it is not conventionally regarded as a major class feature. In phonology, this major class feature model is important to discuss the syllable structure. According to Sonority Sequencing Principle, segments belonging to classes in upper rows in Table 1 have lower sonority, occupying outer positions in a syllable, while those belonging to classes in lower rows have higher sonority, occupying inner positions in a syllable. Ramus and Mehler [7] state that the perception of consonant classes plays an important role in the recognition of "syllabic rhythm"/"broad phonotactics", and this implies that these feature classes serve as one of the perceptual cues to suprasegmental information.

	[cons]	[approx]	[sonor]	Sonority
Obstruents	+	-	-	Low
Nasals	+	-	+	:
Liquids	+	+	+	:
Glides	-	+	+	High

Table 1: Major class features.

The mean identification rates for each major class and subject group are shown in Figure 4. "Average" in Figure 4 (shown in broken lines) represents the mean identification rates across four major classes. While each feature displays a considerable variation yet to be explained, the average shows that both Japanese and English speakers successfully recognised the major classes: Japanese 66-82% and English 77-90%.

Figure 5 displays the mean identification rates calculated for the individual features, across the four major classes. Observing the identification rates for individual features, it is found that the perception is biased so that sonority tends to be perceived lower. Japanese speakers were likely to mishear [-cons] consonants as [+cons] consonants (i.e., mishear higher-sonority consonants as lower-sonority consonants), [+approx] as [-approx], and [+sonor] as [-sonor], and the degree of bias was large in this order. English speakers showed the similar tendency with the rates being a little different.

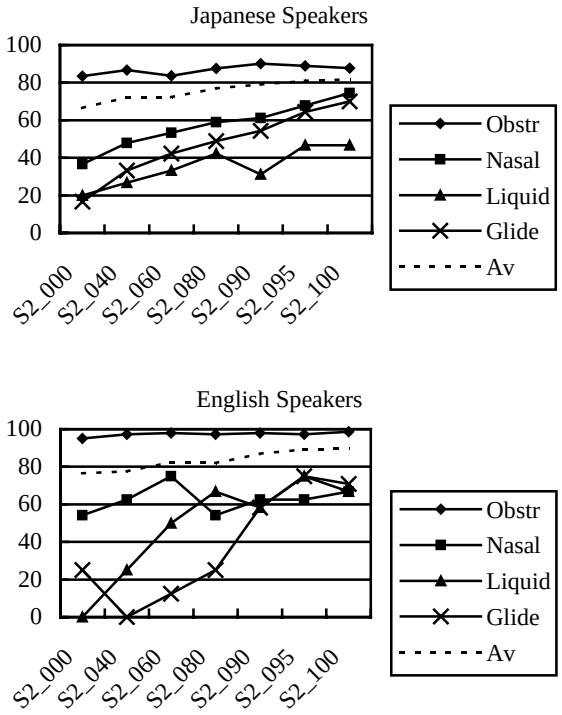


Figure 4: Percentage correct identification of major classes.

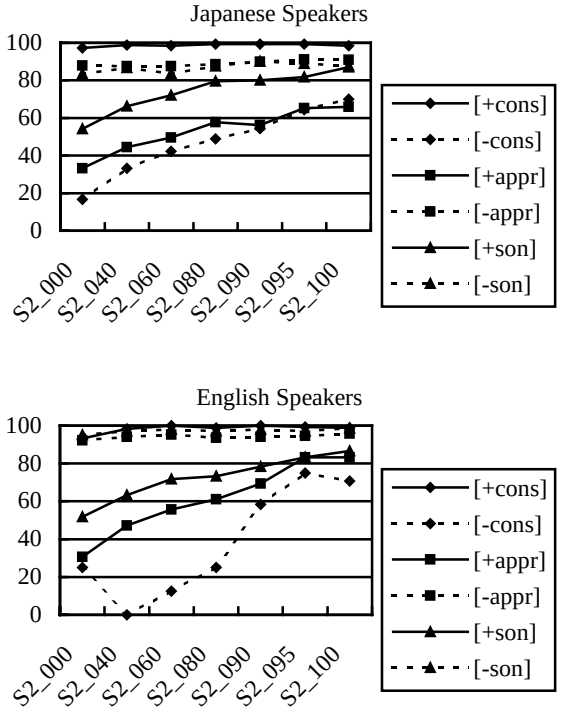


Figure 5: Percentage correct identification of major class features.

4. CONCLUSION

The results of the two experiments suggest that listeners could partially recognise Japanese consonants in LPC residuals, although it was well above the chance level, and that the identification rate rose sharply if one resonance peak was added to those signals. This was also the case with L2 listeners. The phonological analysis of the results indicates that (1) the voice feature was sufficiently identified even in residuals; (2) the identification rates of the major classes were also high in the residuals and the rates rose even higher as the spectral peak became eminent.

5. REFERENCES

1. Mori, K., Toba, N., Harada, T., Arai, T., Komatsu, M., Aoyagi, M., and Murahara, Y. "Human Language Identification with Reduced Spectral Information," *Proceedings of Eurospeech '99*: 391-394, 1999.

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

3 Takeda, K., Sagisaka, Y., Katagiri, S., Abe, M., and Kuwabara, H. *Speech Database User's Manual*. Advanced Telecommunications Research Institute International, 1988.

4 Inozuka, H., and Inozuka, E. *Nihongo no Onsei Nyumon [Introduction to Japanese Phonetics]*. Babel Press, Tokyo, 1933.

5 O'Shaughnessy, D. *Speech Communication: Human and Machine*. Addison-Wesley, Reading, MA, 1990.

6 Kenstowicz, M. *Phonology in Generative Grammar*. Blackwell, Cambridge, MA, 1994.

7 Ramus, F., and Mehler, J. "Language Identification with Suprasegmental Cues: A Study Based on Speech Resynthesis." *J. Acoust. Soc. Am. 105*: 512-521, 1999.

Acknowledgements. This research is funded by The Foundation Hattori-Hokokai. We are grateful to Mr Kenji Sasaki and Ms Aki Osawa of Sophia University for organising the experiments for us.