Modelling the Perceptual Identification of Japanese Consonants from LPC Cepstral Distances

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

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

Abstract

This study attempts to account for the perceptual phenomenon observed in Komatsu et al. [1] in terms of the spectral properties of the LPC re-synthesised stimuli. To implement this, LPC cepstral distances between re-synthesised samples and their original samples are measured. The results of the acoustic analysis and their comparison with the perceptual data indicate that there is a striking similarity in patterns between the spectral property of the Japanese consonants and their perceptual scores. This suggests that the role played by spectral information in the perception of Japanese consonants is significant across all consonant types, and also implies that even in its crudest form, it contributes significantly to their perception.

1. Introduction

Our previous study, Komatsu et al. [1], investigated how spectral information contributes to the perception of Japanese consonants, using re-synthesised samples that were created by (i) gradually reducing the order of LPC analysis in the residual excited LPC vocoder; and (ii) gradually flattening the spectral peak in the frequency domain. The actual stimuli used in the two perceptual experiments were 17 Japanese /C/+/a/ syllables with consonants /p b t d k g s S h tS dZ m n z * j w/. Their spectral information was manipulated as follows:

- Stimuli used in Experiment 1: They 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. This set of 5 types of stimuli, plus the original (unmodified) syllables and the 22nd residuals, were used for Experiment 1. Henceforth this set is called Set 1.
- Stimuli used in Experiment 2: They were created by calculating their 2nd order LPC coefficients and applying them to the 22nd LPC residuals, but in their LPC re-synthesis, the spectral peak of the re-synthesised syllables were changed by multiplying the magnitudes of the poles on the *z*-plane by one of 7 factors (1.00 / 0.95 / 0.90 / 0.80 / 0.60 / 0.40 / 0.00). This set of 7 types of stimuli was used for Experiment 2. Henceforth this set is called Set 2.

The results of native Japanese speakers in Experiment 1 showed that they could partially recognise Japanese consonants in the residuals (average 31%) and that the average identification rate rose sharply in re-synthesised samples with the 2nd order LPC coefficients (average 89%), as reproduced in Figure 1. In Figure 1, the vertical axis represents mean identification rates and the horizontal axis stimulus types in Set 1. "E" represents residuals, "X" the original samples and "S*n*" re-synthesised samples with the *n*th order LPC coefficients. (Henceforth, each stimulus type is called *E*, *S2*, *S4*, *S6*, and so on.)

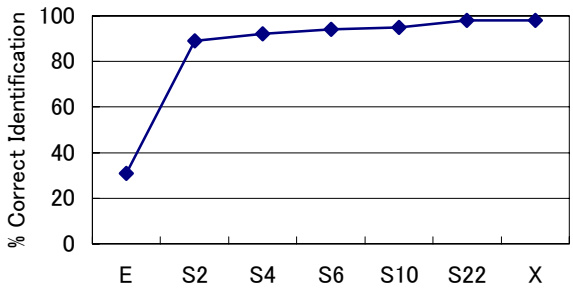

Figure 1: The results of Komatsu et al. [1]; Experiment 1

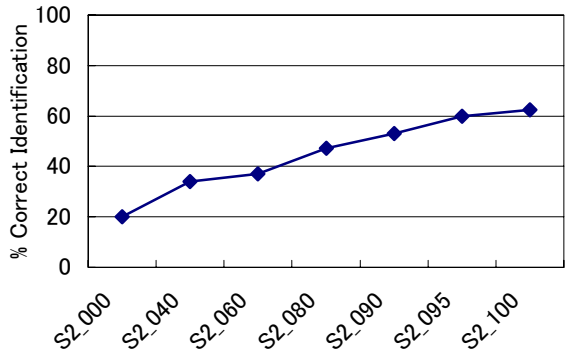

Figure 2: The results of Komatsu et al. [1]; Experiment 2.

The results of Experiment 2, reproduced in Figure 2, showed that the identification rate by native Japanese speakers improved gradually as the factor increased (Note that increasing/reducing the factor sharpens/flattens the spectrum). In Figure 2, the mean consonant identification rates of Experiment 2 for each stimulus type are plotted, and "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).

However, the robustness of consonant perception against spectral reduction, observed in Figure 1, was not thoroughly discussed in Komatsu et al. [1]. In particular, it was not explained why, as seen in Figure 1, there was a sharp rise in identification rates between *E* and *S2*. Furthermore, the spectral properties of the stimuli were not properly analysed to verify that the spectral information was reduced in a consistent way. Consequently, the observed perceptual phenomenon was explained only in terms of the LPC orders, but not in terms of the actual acoustic properties of the stimuli.

In this study, we aim to investigate whether the perceptual phenomenon observed in Komatsu et al. [1] can be explained by the actual spectral properties of the stimuli. We also aim to study the relationship between the perceptual data and the obtained results of the acoustic analysis. To implement this acoustic analysis, LPC cepstral distances between stimuli are measured. This is because cepstral distance analyses are widely used for automatic speech recognition and for the evaluation of the quality of coded speech [2].

2. Experiment: Acoustic analysis

2.1. Materials

Stimuli Set 1 and Set 2 used in the two perceptual experiments of Komatsu et al. [1]: 17 Japanese /C/+/a/ syllables whose spectral information was manipulated as described in Section 1.

2.2. Procedures

The cepstral distance of each synthesised sample from its corresponding original sample was calculated in the following stages: (i) all the materials used in the experiments were re-analysed with the 22nd order LPC (using Hamming window, frame: 512 points, 75% overlap); (ii) cepstrum was calculated from the obtained LPC coefficients; (iii) using the first 22 cepstral coefficients, the cepstral distance between the original sample (= *X*) and the synthesised sample was calculated at each frame; and (iv) the cepstral distances from each frame were pooled across the consonant section to obtain the mean distance for the section, which was determined by inspecting waveforms and wide-band spectrograms of the original sample.

2.3. Results

For each Set 1 stimulus type (i.e. *E*, *S2*, *S4*, etc.) except the original, the obtained LPC cepstral distances were pooled across all 17 consonant types. Then their means were obtained and they are plotted in Figure 3. The vertical axis in Figure 3 shows the distance in decibels, and the horizontal axis represents the Set 1 stimulus type. "E" represents the residuals. Since the distances are measured from the original syllables, they are inversely related to how similar the synthesised syllables are to the original syllables, i.e. the larger the distances are, the more spectrally distorted the stimuli are from the originals. Figure 3 shows the following:

- Overall cepstral distances increase as the order of LPC coefficient decreases, which confirms that the spectral information of the Set 1 stimuli was modified as we intended in the perceptual experiment.
- There is a conspicuous increase in LPC cepstral distance between *E* (= 22nd order LPC residuals) and *S2* (= re-synthesised samples with the 2nd order LPC coefficient). This suggests that the sharp rise in the identification rate between *E* and *S2*, as shown in Figure 1, can be ascribed to the significant change in LPC cepstral distances between *E* and *S2*.

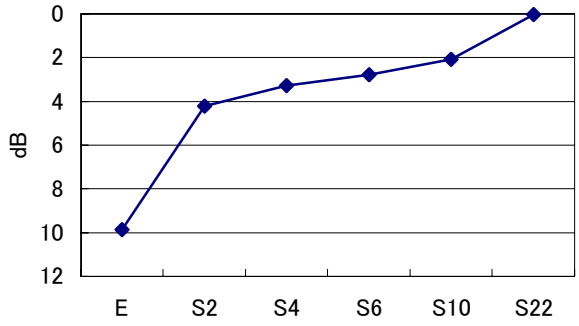

Figure 3: Mean LPC cepstral distance; Set 1

It must be mentioned that these two tendencies were also observed in the mean cepstral distances obtained for the consonant section plus the consonant-to-vowel transition, or for the entire /CV/ syllable, although figures are not produced here due to the limited space.

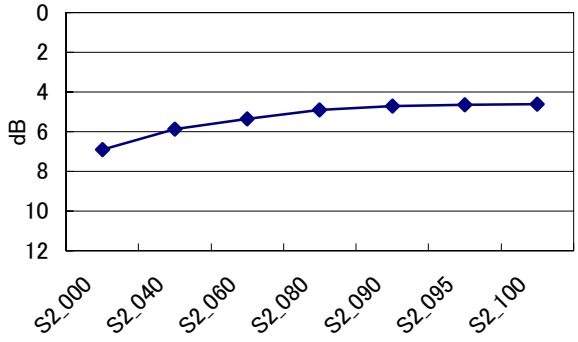

Figure 4: Mean LPC cepstral distance; Set 2

Figure 4 shows the mean LPC cepstral distances calculated across all 17 consonant types for each Set 2 stimulus type. As in Figure 3, the vertical axis shows the distance in decibels, and the horizontal axis the Set 2 stimulus type. It indicates that as in Figure 3, overall cepstral distances increase as the peak of the LPC spectral peak reduces. This tendency is certainly the mirror image of the results of the perceptual Experiment 2 shown in Figure 2, and it proves that in perceptual Experiment 2, the spectral information was manipulated consistently.

2.4. Discussion

One of the objectives of this study is to investigate why there was a sharp rise in identification rates between *E* and *S2*. The results of the acoustic analysis demonstrated that there is an

unmistakable decrease in the spectral distance from *E* to *S2*, and it is assumed that this caused the sudden rise of identification rate there. LPC analysis separates speech signals into residuals (representing source information) and coefficients (representing filter information), and hence the filter used to re-synthesise *S2* had only one pole. This means that *S2* contains minimal spectral information. However, this slight addition of spectral information dramatically improved the spectral quality of the consonant, and consequently it enhanced the identification rate significantly, regardless of consonant types.

It is also assumed that the addition of spectral information does not greatly improve the perception score from *S2* to *S22* because the spectral information provided by the signals was 'sufficient': the improvement of the spectral quality became 'redundant' after a certain level, although it contributed to the rise in identification rate to a certain extent.

Now there is one potential criticism that needs to be addressed on the results of the experiment: although the mean identification scores (Figures 1 and 2) and the mean LPC cepstral distances (Figures 3 and 4) show the similar pattern, this may be the product of averaging: identification scores and cepstral distances may differ from consonant to consonant, displaying inconsistent patterns.

To address this criticism, the mean LPC cepstral distances of Set 1 stimuli were obtained for each LPC analysis order and consonant type, and they are plotted in Figure 5. Also the mean identification scores of Set 1 stimuli were calculated for each consonant type. They are shown in Figure 6. In Figures 5 and 6, the consonants are arranged in the order of identification scores of residuals (= *E*).

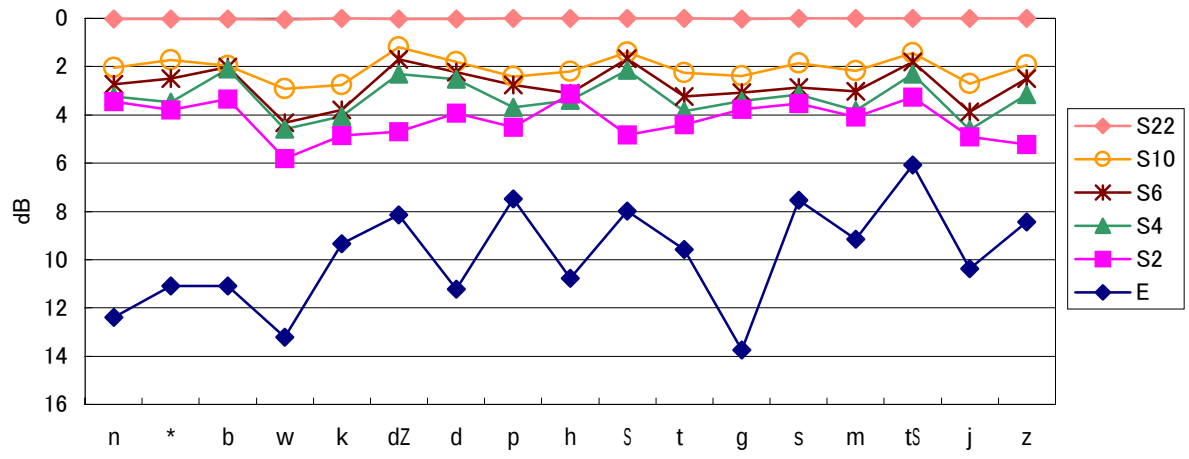

Figure 5: Mean cepstral distance for each consonant

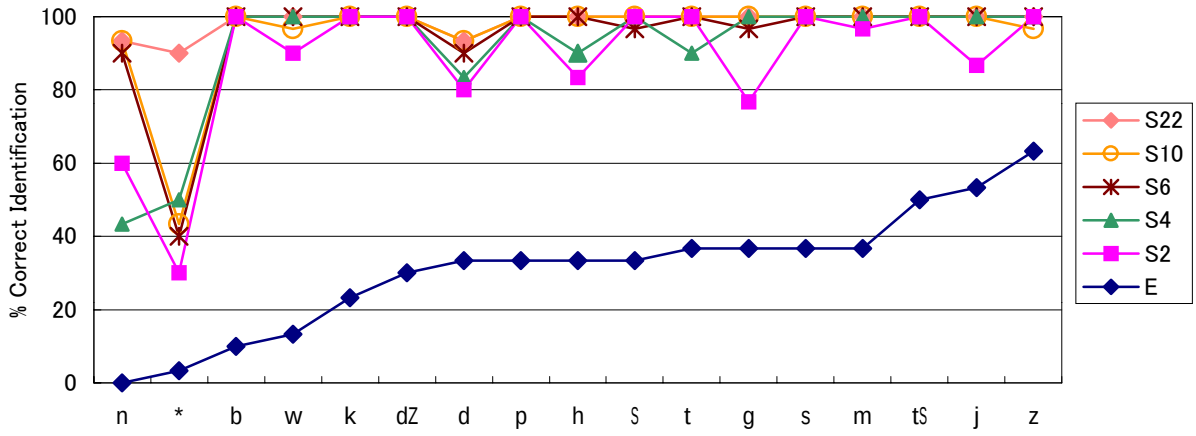

Figure 6: Mean identification scores for each consonant

Figures 5 and 6 confirm that the observations which were made with regard to the average scores and distances hold for all consonants: there are sharp rises in LPC cepstral distances (Figure 5) and identification rates (Figure 6) between *E* and *S2* across all consonants. Figure 5 also demonstrates that the spectral distances across all consonant types increase as the number of LPC coefficients reduces.

One may claim that the identification rate of /*/ does not show a sharp rise in Figure 6, but this absence of the sharp rise does not weaken the overall trend, since there is certainly a significant, if not large, gap in the identification rate of /*/ between *E* and *S2*. In fact, there is a gap in perceptual scores of /*/ between *S10* and *S22*, and this can be attributed to the gap in the spectral distance between *S10* and *S22*, where the spectral information provided by the stimuli was sufficient for the other consonants.

It should be noticed that in Figure 5, the LPC cepstral distance of *E* tends to decrease from left to right. This means that, since the consonants are arranged in the order of the identification rate in Figures 5 and 6, the LPC cepstral distance of *E* decreases across all consonants as the identification rate rises. The correlation coefficient *r* calculated between the cepstral distances and the identification scores for *E* is -0.51, showing an inverse relationship between them. This provides another supportive piece of evidence for the close relationship between spectral distance and the perceptual score.

To summarise, the comparison between the results of the acoustic analysis of the stimuli used in Komatsu et al. [1] and their perceptual scores indicates that there is a striking similarity in patterns between the spectral property of the Japanese consonants and their perception scores, and this implies that the former contributes significantly to the latter.

3. Implication from the obtained results

A common practice in traditional consonant perception experiments is to test a hypothesis about the perceptual significance of certain acoustic characteristics of signals, and in this way it has been found that the release burst is important for plosives, high frequency spectral patterns for fricatives, and so on. However, it has been difficult to assign a single coherent set of acoustic measures to describe all of the consonants. This is because consonants differ significantly among themselves in their acoustic properties (see, for example [3]). Because of these differences, the perception of consonants cannot be modelled as a whole in terms of a simple universal acoustic parameter, like duration or formant patterns in the case of vowel perception. Instead, it is usually compelled to be studied in the subsets of the whole consonant inventory, such as plosives, fricatives, or in minimal pair contrasts, such as /S/-/tS/. The weakness of an approach only based on a small set of consonants is that it may miss the simple and robust features of sounds that characterise the speech processing as a whole.

From the obtained results, one can argue that, regardless of the consonant type, the spectral information, even in a crudest form (= *S2*), contributes significantly and universally to the perception of all Japanese consonants. This is a step towards a unified approach to consonant perception; this could provide an alternative to a conventional approach that pursues perceptual models which show great complexity in cue interaction associated with each consonant category. (This unified approach to consonant perception is also discussed in Choo [4][5], which extended the spectral-distance analysis of vowels to fricatives.) However, it must be emphasised that the issue to choose between the two approaches is beyond the scope of this paper, and it should be dealt with in a separate study.

4. Conclusion

The results of the acoustic analysis and their comparison with the perceptual data indicate that there is a striking similarity in patterns between the spectral property of the Japanese consonants and their perception scores. This suggests that the role played by spectral information in perception of Japanese consonants is significant and robust regardless of the consonant type, and also implies that even in its crudest form, it contributes significantly to their perception.

5. Acknowledgements

This work is partially supported by Special Research Grant from Sagami Women's University and by 1999 grant from The Foundation Hattori-Hokokai.

6. References

[1] Komatsu, M., Tokuma, W., Tokuma, S., and Arai, T., "The effect of reduced spectral information on Japanese consonant perception: Comparison between L1 and L2 listeners," *Proceedings of the 6th International Congress on Spoken Language Processing (ICSLP 2000), Beijing, China*, Vol. 3, 750-753, 2000.

[2] Furui S., *Digital Speech Processing, Synthesis, and Recognition*, 2nd ed., Marcel Dekker, NY, 2001.

[3] Kent, R. D. and Read, C., *The Acoustic Analysis of Speech*, Singular Publishing Group, San Diego, 1992.

[4] Choo, W., *Relationships between Phonetic, Perceptual and Auditory Spaces for Fricatives*, PhD Dissertation, University of London, 1996.

[5] Choo, W., "The relationship between perceptual and physical space of fricatives," *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS 99), San Francisco, California*, 163-166, 1999.