PERCEPTUAL ROLES OF DIFFERENT FREQUENCY BANDS IN JAPANESE SYLLABLE IDENTIFICATION

Kazuo Ueda and Yoshitaka Nakajima
Department of Human Science and Center for Applied Perceptual Research
Kyushu University, 4-9-1 Shiobaru Minami-ku, Fukuoka 815-8540, Japan
Kiyoto Noguchi and Yuichi Satsukawa
Department of Acoustic Design, Kyushu University, 4-9-1 Shiobaru Minami-ku
Fukuoka 815-8540, Japan
ueda@design.kyushu-u.ac.jp nakajima@design.kyushu-u.ac.jp

Abstract

Ueda et al. [(2010). Fechner Day 2010, Padua.] indicated that speech information could be essentially transmitted by the power fluctuations in four frequency bands. We aimed at clarifying the roles of these frequency bands in Japanese speech perception in V/CV syllable identification. We first performed factor analyses of power fluctuations of critical-band-filtered speech, and obtained four frequency bands as in the previous research. The speech was a set of V/CV patterns uttered by a male and a female speaker. The speech patterns were converted into noise-vocoded speech so that only the power fluctuation in each frequency band was preserved. There were also patterns in which one of the frequency bands was eliminated resulting in a spectral gap. Eliminating the lowest band (50-570 Hz) crucially deteriorated perceptual differentiation between voiced and unvoiced consonants. Eliminating the second lowest band (570-1850 Hz) interfered vowel identification turning almost all vowels into /i/. The roles of the other frequency bands were not obvious, but their temporal relationships with the lowest band was suggested to play a role.

The aim of the present investigation was to clarify how elimination of one of four frequency-bands in noise-vocoded speech of Japanese V/CV syllables affected identification, in terms of the amount of information transmitted obtained from confusion matrices.

Ueda, Nakajima, and Satsukawa (2010) extracted three factors of power fluctuations common to eight languages. These factors had four non-overlapping mounds in factor loadings, and the frequency range of speech were divided into four frequency-bands separated by boundaries around 500, 1700, and 3300 Hz; each band centered around one of the mounds. Nakajima, Ueda, Fujimaru, Motomura, and Ohsaka (2012) examined the correspondence between factor scores and phonemic labels in British English, and found that these three factors clearly differentiated three phonemic categories: vowels, sonorant consonants, and obstruents. These factors were correlated to *sonority* or *aperture* (e.g., de Saussure, 1959; Selkirk, 1984; Spencer, 1996).

Another way to explore potential correspondence between those factors and phonemic perception would be to simplify the three factors into power fluctuations in the four frequency bands by utilizing noise-vocoded speech (e.g., Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995; Smith, Delgutte, & Oxenham, 2002; Sheldon, Pichora-Fuller, & Schneider, 2008; Roberts, Summers, & Bailey, 2011), and observing the effects of manipulating these frequency bands on perception. Miller and Nicely (1955) and Benkí (2003) estimated the amount of information transmitted through human listeners, based on confusion matrices of consonant identification. Employing a similar analysis method,

Table 1 Experimental conditions.

Condition	Number of bands	Eliminated band (Hz)
1: Original	-	-
2: 20-band noise-vocoded	20	-
3: 4-band noise-vocoded	4	-
4: 1st-band eliminated	3	50-570
5: 2nd-band eliminated	3	570-1850
6: 3rd-band eliminated	3	1850-4000
7: 4th-band eliminated	3	4000-7000

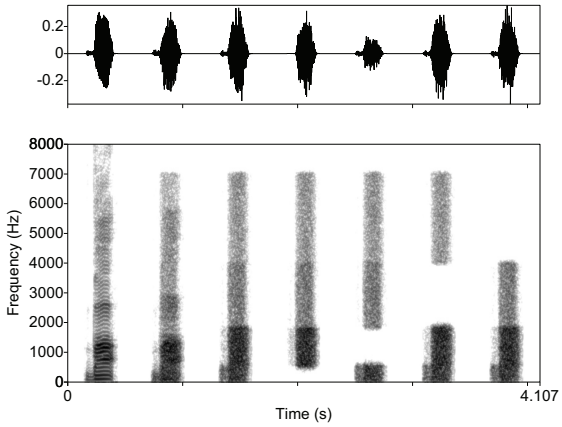

Figure 1 Examples of stimuli based on a Japanese CV-syllable, /ba/. The upper panel shows the waveforms and the lower panel the corresponding spectrograms. In each panel, from left to right, the experimental conditions 1 to 7 in Table 1 are exemplified.

we eliminated one of the frequency-bands in noise-vocoded speech of Japanese V/CV syllables, and estimated the amount of information transmitted by each frequency band.

Experiment

Method

Experimental conditions and stimuli Experimental conditions are listed in Table 1. A total of 101 Japanese V/CV syllables uttered by a male and a female speaker was edited and extracted from "ATR Digital Speech Database Set A" (20-kHz sampling and 16-bit quantization). In condition 1, the edited recordings of Japanese V/CV-syllables were played. Noise-vocoded speech stimuli were prepared for conditions 2-7. Critical bands (e.g., Zwicker & Terhardt, 1980) were adopted over the range of 50-7000 Hz in condition 2. Examples of stimuli are shown in Figure 1.

Table 2 Japanese consonant classification utilized in the analysis.

Consonant classification	Phonemes
Voicing	
Voiced	/b, d, g, z, m, n, N, ɾ, y, w/
Unvoiced	/p, t, k, s, h/
Manner of articulation	
Stops	/p, t, k, b, d, g/
Fricatives	/s, z, h/
Nasals	/m, n, N/
Flap	/ɾ/
Semivowels	/y, w/

Participants All the participants were normal-hearing Japanese speakers. Eight listeners (5 males and 3 females, mean age = 26, ranged from 21 to 45) participated in experiment sessions where stimuli based on syllables uttered by a male speaker were presented. Ten other listeners (5 males and 5 females, mean age = 23, ranged from 22 to 25) participated in experiment sessions where stimuli based on syllables uttered by a female speaker were presented.

Procedure The stimuli were diotically presented to each listener through headphones (STAX SR-303). Each stimulus allotted to three trials in random order for each listener. Within each trial, presentation of the same stimulus was repeated three times in succession with an inter-stimulus-interval of about 2 s. The sound pressure level was adjusted so that the peak level of a vowel /a/ became 78 dB A (Fast). The listeners were instructed to identify what they heard by selecting an appropriate button on a screen as far as possible. When they could not identify a stimulus, they were instructed to press a "?" button, or to write down what they heard on an answer sheet.

Results

The listeners' responses were collected as confusion matrices of 101×101. The responses that fell outside the matrices were neglected, because they were rare (0.08% in the highest condition). Relative amount of information transmitted was calculated according to the method utilized by Benkí (2003). Separate information channels were assumed for vowels and consonants in the calculation. Vowels were classified into five categories, /a, e, i, o, u/. The classification utilized for consonants was summarized in Table 2. Figures 2 and 3 show the relative amounts of information transmitted. Figure 2 shows three effects on the relative amounts of information transmitted: the effects of noise-vocoding, number of bands, and elimination of specific frequency bands. Noise-vocoding generally reduced the amounts of information transmitted, except in the 20-band condition in vowel distinction (Fig. 2a). Reducing the number of bands from 20 to 4 resulted in reducing the amounts of information transmitted, except in voicing distinction (Fig 2b). Eliminating a specific frequency band, especially the second lowest band in vowel distinction and the lowest band in voicing distinction, largely reduced the amounts of information transmitted (Fig. 2a,b). Figure 3 shows similar effects concerning noise-vocoding and the number

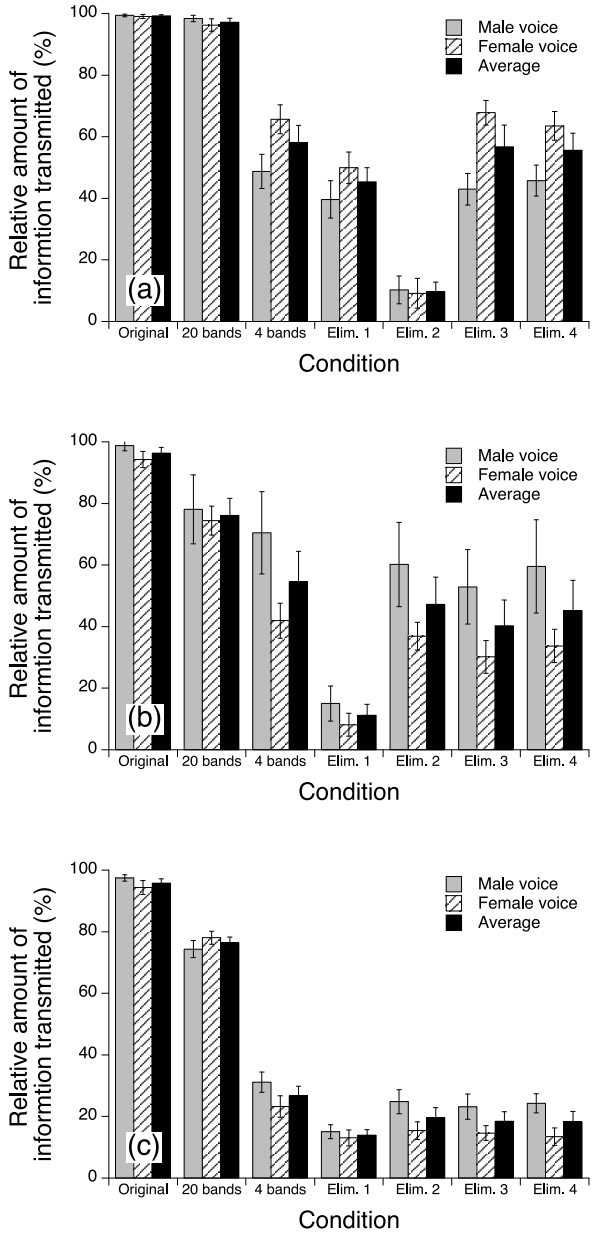

Figure 2 Relative amounts of information transmitted, calculated in terms of identifying (a) vowels, (b) voicing, and (c) manners of articulation. The error bars represent 95% confidence intervals.

of bands, whereas the effect of eliminating specific frequency bands differed very much between consonant classification categories. The amounts of information transmitted for fricative distinction reduced sharply when the lowest frequency band was eliminated (Fig. 3b). A nonparametric multiple comparison test (Steel-Dwass test) supported these

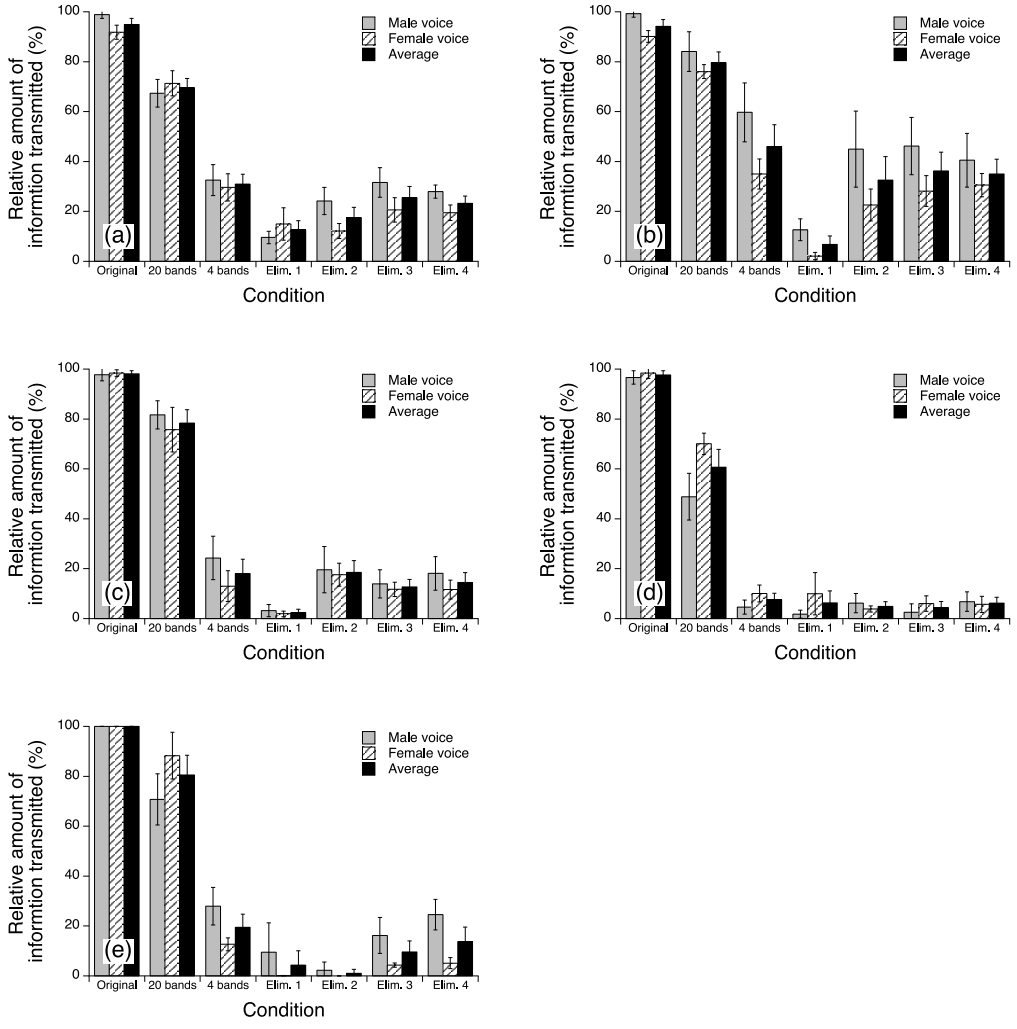

Figure 3 Relative amounts of information transmitted, calculated in terms of identifying categories within manners of articulation: (a) stops, (b) fricatives, (c) nasals, (d) flap, and (e) semivowels. The error bars represent 95% confidence intervals.

observations.

Discussion

The lowest frequency band should not be removed when information concerning voicing of consonants should be conveyed. It is possible that a short forerunning energy in the lowest frequency band preceding a burst, which can be seen in Figure 1, signals a voiced consonant, as a substitute of a voice bar, whereas the absence of the lowest frequency band tends to signal a voiceless consonant. Thus, voicing can be cued without periodicity of voicing. Removing the second lowest frequency band may distort formant structure of vowels, resulting in a fixed formant pattern which resembles that of the vowel /i/.

The roles of the third and the fourth frequency bands were not obvious in the present investigation, but, it is possible that these frequency bands have some roles in phoneme perception in combination with other frequency bands; the amounts of information transmitted to signal fricatives may be related to the temporal relationship between the lowest frequency band and the third or fourth band.

Acknowledgments

This research was supported by Grants-in-Aid for Scientific Research Nos. 14101001, 19103003, and 20330152 from the Japan Society for the Promotion of Science, Kyushu University Interdisciplinary Programs in Education and Projects in Research Development, and a Grant-in-Aid for the 21st Century COE program from the Ministry of Education, Culture, Sports, Science and Technology.

References

Benkí, J. R. (2003). Analysis of English nonsense syllable recognition in noise. *Phonetica*, *60*, 129-157.

de Saussure, F. (1959). *Course in general linguistics (Baskin, W., Trans.)*. New York: McGraw-Hill Paperbacks. (Original work published 1916)

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.*, *27*, 338-352.

Nakajima, Y., Ueda, K., Fujimaru, S., Motomura, H., & Ohsaka, Y. (2012). Acoustic correlate of phonological sonority in British English. In *Fechner Day 2012, Proceedings of the 28th Annual Meeting of the International Society for Psychophysics.* Ottawa, Canada.

Roberts, B., Summers, R. J., & Bailey, P. J. (2011). The intelligibility of noise-vocoded speech: spectral information available from across-channel comparison of amplitude envelopes. *Proc. Royal Soc. B*, *278*, 1595-1600.

Selkirk, E. (1984). On the major class features and syllable theory. In M. Aronoff & R. T. Oehrle (Eds.), *Language Sound Structure: Studies in Phonology Presented to Morris Halle by His Teacher and Students* (p. 107-136). Cambridge, MA: MIT Press.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303-304.

Sheldon, S., Pichora-Fuller, M. K., & Schneider, B. A. (2008). Effect of age, presentation method, and learning on identification of noise-vocoded words. *J. Acoust. Soc. Am.*, *123*, 476-488.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(7 March 2002), 87-90.

Spencer, A. (1996). *Phonology: Theory and Description*. Oxford: Blackwell.

Ueda, K., Nakajima, Y., & Satsukawa, Y. (2010). Effects of frequency-band elimination on syllable identification of Japanese noise-vocoded speech: Analysis of confusion matrices. In *Fechner Day 2010, Proceedings of the 26th Annual Meeting of the International Society for Psychophysics* (p. 39-44). Padova, Italy.

Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.*, *68*, 1523-1525.