AN ACOUSTIC LANGUAGE UNIVERSAL: PERCEPTUAL EXPERIMENTS EMPLOYING NOISE-VOCODED SPEECH

Kazuo Ueda and Yoshitaka Nakajima
Department of Human Science, Kyushu University, 4-9-1 Shiobaru, Minami-ku, 815-8540 Fukuoka, Japan
ueda@design.kyushu-u.ac.jp nakajima@design.kyushu-u.ac.jp

Tomoya Araki
Unit of Perceptual Psychology, Kyushu University, 4-9-1 Shiobaru, Minami-ku, 815-8540 Fukuoka, Japan
araki@gsd.design.kyushu-u.ac.jp

Abstract

A consistent clustering of frequency bands in different languages and dialects, i.e., English, French, German, Japanese, Mandarin, and Cantonese, had been found through factor analyses applied to power fluctuations of critical-band filtered speech sounds [Ueda & Nakajima, Trans. Tech. Comm. Psychol. Physiol. Acoust., 38, 771-776, (2008); 39, 211-216, (2009)]. Four frequency bands appeared commonly to these languages and dialects. The perceptual importance of these frequency bands were, however, largely unknown. We report a series of perceptual experiments, in which we employed Japanese noise-vocoded speech that was synthesized to reconstruct amplitude envelope patterns in some frequency bands of original speech. Almost perfect sentence recognition was achieved without training by using noise-vocoded speech synthesized with the four frequency bands above mentioned. The relative importance of power fluctuations in low frequency bands was revealed. Spectral level patterns averaged in time had a small but significant effect on the intelligibility. These results shed lights on the critical roles, in speech perception, of the power fluctuations in low frequency bands.

The present investigation focuses on how frequency bands, in which power fluctuations exist, are critical to the intelligibility of noise-vocoded speech.

Ueda and Nakajima (2008a,b; 2009) found that speech information of different languages and dialects, namely, British English, French, German, Japanese, Mandarin, and Cantonese, can be represented through four common frequency bands with power fluctuations without a temporal fine structure of speech. They filtered 200 sentences of each of British English, French, German, and Japanese, 78 sentences of Mandarin, and 58 sentences of Cantonese, each uttered by 10 native speakers (5 females and 5 males), through a bank of 20 bandpass filters, i.e., critical-band (Fletcher, 1940; Zwicker and Terhardt, 1980) filters. Power fluctuations calculated in each band yielded matrices of correlation coefficients across critical bands. These matrices were subject to factor analyses to find clusters of critical bands. The same three factors, through which four clusters of critical bands were represented, were consistently obtained over all languages examined. The four clusters represented four frequency bands, and we could calculate crossover frequencies of the curves as boundaries of these frequency bands. The frequency boundaries were very consistent across different languages. We could synthesize intelligible enough noise-vocoded speech utilizing these boundary frequencies.

One of the factors exhibits a characteristic shape with two peaks. This implies some degree of correlation between the lowest and the second highest frequency band. Then, we hypothesized that a deteriorative effect of blending the amplitude envelopes of these two bands should be small compared with the other blending combinations of frequency bands. Thus, the purpose of Experiment 1 was to test this hypothesis on the intelligibility of noise-vocoded speech stimuli blending amplitude envelopes between frequency bands.

Experiment 1

Method

Stimuli and conditions A total of 85 Japanese sentences uttered by a male native speaker were extracted from a commercial speech database (NTT-AT, "Multi-lingual speech database 2002"). These speech sentences had been recorded with 16-kHz sampling and 16-bit quantization. The recorded speech samples were edited to eliminate unnecessary blanks and noise, leaving silent margins of about 10 ms before and after each sentence.

Noise-vocoded speech was synthesized from those edited speech samples. The speech samples passed through either a 20- or a 4-channel filter bank (Table 1), and then the amplitude envelopes were extracted at each channel output. The envelopes were calculated as the following. Each filter output was squared to obtain power. The power was smoothed by a moving average of a 15-ms Gaussian window. The power ratio between the filtered speech and a bandnoise with a corresponding bandwidth was obtained. The bandnoise was modulated with the root of the ratio, which corresponded to an amplitude envelope. Some of the amplitude envelopes were blended in the dimension of squared amplitude, according to the

Table 1. Filter settings. The passbands are expressed in Hz. The 20-band setting was based on critical bandwidths (Zwicker and Terhardt, 1980). The 4-band setting was based on the frequency boundaries found by Ueda and Nakajima (2008a, 2008b, and 2009).

20 bands		4 bands	
Band number	Passband	Band number	Passband
1	50-150		
2	150-250		
3	250-350	1	50-570
4	350-450		
5	450-570		
6	570-700		
7	700-840		
8	840-1 000		
9	1 000-1 170	2	570-1 850
10	1 170-1 370		
11	1 370-1 600		
12	1 600-1 850		
13	1 850-2 150		
14	2 150-2 500		
15	2 500-2 900	3	1 850-4 000
16	2 900-3 400		
17	3 400-4 000		
18	4 000-4 800		
19	4 800-5 800	4	4 000-7 000
20	5 800-7 000		

Table 2. Experimental conditions. Blended bands are in parentheses.

Condition	Number of bands in analysis	Number of bands in synthesis	Band blending condition
1	20	20	1, 2, 3,..., 20
2	20	4	(1-5), (6-12), (13-17), (18-20)
3	4	4	1, 2, 3, 4
4	4	3	(1, 2), 3, 4
5	4	3	(1, 3), 2, 4
6	4	3	(1, 4), 2, 3
7	4	3	1, (2, 3), 4
8	4	3	1, (2, 4), 3
9	4	3	1, 2, (3, 4)
10	4	2	(1, 2), (3, 4)
11	4	2	(1, 3), (2, 4)
12	4	2	(1, 4), (2, 3)
13	4	2	(1, 2, 3), 4
14	4	2	(1, 2, 4), 3
15	4	2	(1, 3, 4), 2
16	4	2	1, (2, 3, 4)
17	4	1	(1, 2, 3, 4)

experimental conditions (Table 2). Modulated bandnoises were added up to form noise-vocoded speech. Eighty-five sentences were randomly assigned to one of the 17 conditions for each participant. As a consequence, five sentences were assigned to each condition.

Procedure Five sessions, each containing 17 trials of different conditions, were formed for each participant. Stimuli were presented to both ears of the participant in a soundproof booth through a computer (HP Compaq, 6710b), a USB audio adaptor (Roland, UA-4FX), a headphone adaptor (STAX, SRM-1/MK-2), and headphones (STAX, SR-303). The average presentation level was adjusted to about 74 dB SPL, using a 1-kHz calibration tone included in the speech database. The sound pressure level was measured using an IEC coupler (Brüel & Kjær, 4153), a microphone (Brüel & Kjær, 4192), and a precision sound level meter (Brüel & Kjær, 2260).

Each sentence in a trial was presented three times in succession to the participants with an inter-stimulus-interval of about 1 s. Fourteen native speakers of Japanese (five females and nine males) with normal hearing participated. The participants were instructed to write down what they had heard without guessing. When they could not hear out a whole sentence but could hear some fragments of a sentence, they estimated where the positions of the fragments were along a guideline printed on an answer sheet, of which the length represented the duration of a sentence. No feedback or systematic training was given to the participants during the whole sessions.

Results

Figure 1 represents average mora accuracy based on the performance in the last four sessions of each participant. We took mora accuracy because any mora can be defined definitely in Japanese. The number of synthesizing bands was the most prominent factor determining the performances [$F(4, 52) = 434.3$, $p < 0.05$]. Closer examination with a multiple comparison (the Tukey's HSD test, $p < 0.05$) revealed that the deteriorating effects came from the conditions of 3-bands or less. Within 3-band-synthesis conditions (conditions 4-9), condition 4 vs. conditions 8 and 9 yielded significant differences. Within 2-band-synthesis conditions

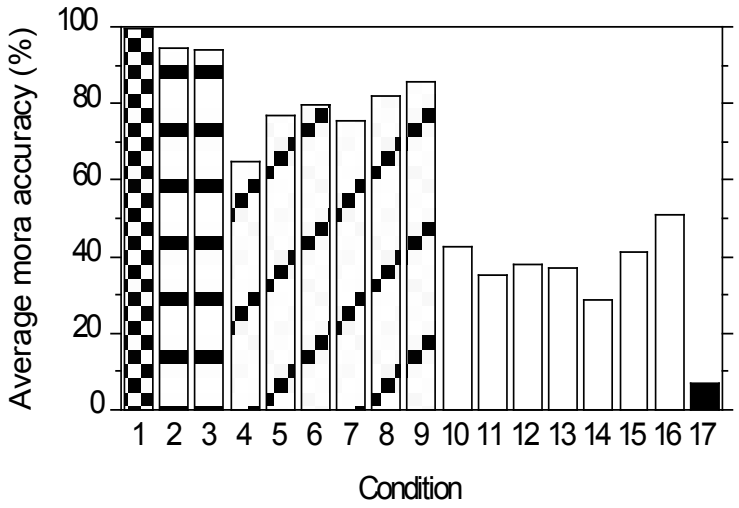

Fig. 1. Average mora accuracy in Experiment 1. The patterns filled in the bars represent numbers of band(s) in synthesis (see Table 2 for detail). The number of band(s) in synthesis firstly determined the performances.

(conditions 10-16), condition 14 vs. condition 16 yielded a significant difference. The differences between condition 17 vs. all the other conditions were significant.

Discussion

Nearly perfect performance was obtained in the 20- and 4-band-synthesis conditions (conditions 1 through 3). The performance in the 4-band-synthesis conditions (conditions 2 and 3) was remarkable, given that no feedback or training was provided to the participants; previous investigations indicated an extensive training was necessary to achieve more than 90% word accuracy (Shannon et al., 1995; Sheldon, Pichora-Fuller, and Schneider, 2008; Davis et al., 2005). The high performance without training exhibited in the present study may be partly explained by the way of determining frequency settings in noise vocoding. The other possible cause may be repetition introduced during stimulus presentation.

Blending amplitude envelopes of two or more frequency bands caused a reduction in the number of independent frequency bands. The effects of blending bands were not so simple as we had expected. This calls for a more systematic control of variables. If we would like to keep a number of independent frequency bands and to examine how independent those frequency bands are, exchanging amplitude envelopes of two frequency bands should be an option. Dorman, Loizou, and Rainey (1997) reported deteriorating effects on intelligibility of noise-vocoded speech, when the amplitude envelopes of four frequency bands were exchanged. However, whether the effects were caused by a particular combination of exchanging frequency bands was unclear, because they exchanged the envelopes in two combinations of frequency bands at once. Moreover, exchanging amplitude envelopes between two frequency bands could have a confounding effect, i.e., both average spectral levels and power fluctuations were simultaneously exchanged in this case. Thus, we conducted two other experiments to separate the effects of exchanging average spectral level patterns (Experiment 2) and power fluctuations (Experiment 3) on the intelligibility.

Experiment 2

Method

Thirty-five Japanese sentences uttered by a male speaker were extracted from the same database as in Experiment 1. All stimuli were 4-band noise-vocoded speech. The average spectral level in each of the four bands was calculated for each original sentence. The gain of each frequency band was adjusted according to seven conditions: a control condition, in which the original spectral levels were kept, and six experimental conditions, in which the average spectral levels were exchanged in every possible combination of the frequency bands. The assignment of the 35 sentences to the seven conditions was randomly decided for each participant. The other parts of noise-vocoding and the experimental procedure were basically the same as in Experiment 1, except for participants (nine participants consisting of three females and six males) and the average presentation level of stimuli (about 60 dB SPL).

Results and Discussion

Figure 2a shows the results. Arcsine-transformed mora accuracy was submitted to multiple comparison (Tukey), which yielded significant differences ($p < 0.05$) between the conditions of high accuracy, i.e., C, (1, 2), and (1, 3) and the conditions of low accuracy, i.e., (2, 3) and (2, 4), where C refers to the control and the numbers in parentheses represent the frequency bands of exchange. Decreasing the average spectral level of the second lowest frequency band seems to have a small but significant deteriorating effect.

Experiment 3

Method

The experimental method was almost the same as in Experiment 2, except that another 35 Japanese sentences were employed, and power fluctuation patterns were exchanged across frequency bands, whereas an average spectral level in each frequency band was kept. This yielded seven conditions including a control. The same nine participants as in Experiment 2 took part. The order of Experiments 2 and 3 was counter-balanced among the participants.

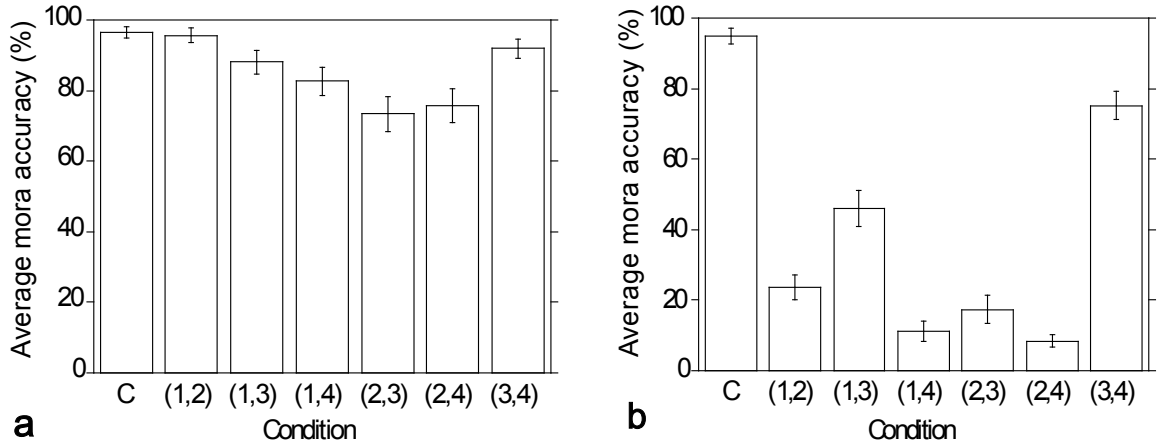

Figure 2. Average mora accuracies in Experiments 2 (**a**, average spectral levels were exchanged) and 3 (**b**, power fluctuation patterns were exchanged). C refers to the control. Numbers in parentheses represent the exchanged frequency bands. Error bars show s.e.m.

Results and Discussion

Figure 2b shows the results. Arcsine-transformed mora accuracy was submitted to multiple comparison (Tukey), which yielded significant differences among C, (1, 3), and (3, 4) conditions ($p < 0.05$). The differences between these three conditions and the other four conditions were also significant. The positions of the two lowest frequency bands (bands 1 and 2) seem to be most critical, probably because these two bands closely related to vowel perception. The combination of bands 1 and 3 had less severe effect, probably because these two bands correlate with each other to some extent (Ueda and Nakajima, 2008a).

General Discussion

The two lowest frequency bands seem to play an essential role in the perception of noise-vocoded speech. Especially, the second lowest band (band 2) should keep both its average spectral level and power fluctuation pattern in the original frequency band to maintain high intelligibility of noise-vocoded speech. It is conceivable that the band 2 acts as a perceptual anchor to the other bands in perceiving speech sounds.

Acknowledgements

This research was supported by Grants-in-Aid for Scientific Research Nos. 14101001, 19103003 and 20330152 from JSPS and by a Grant-in-Aid for the 21st Century COE program from MEXT.

References

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134,* 222-241.

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *Journal of the Acoustical Society of America, 102,* 2403-2411.

Fletcher, H. (1940). Auditory patterns. *Reviews of Modern Physics, 12,* 47-65.

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

Ueda, K., & Nakajima, Y. (2008a). *Factor analyses of critical-band-filtered speech of British English and Japanese.* Paper presented at the Acoustics'08 Paris, Paris, France, *J. Acoust. Soc. Am.*, 123, 3163.

Ueda, K., & Nakajima, Y. (2008b). A consistent clustering of power fluctuations in British English, French, German, and Japanese. *Trans. Tech. Comm. Psychol. Physiol. Acoust., 38*(H-2008-136), 771-776.

Ueda, K., & Nakajima, Y. (2009). Factor analyses of critical-band-filtered speech of Mandarin and Cantonese. *Trans. Tech. Comm. Psychol. Physiol. Acoust., 39*(H-2009-39), 211-216.

Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America,* 68, 1523-1525.