is largest for the baseline condition, followed by the +50 and +80 conditions, respectively.

It is interesting to note that the portion of the Laplace dprime range occupied by the Laplace dprime between stimuli 1 and 2 (25 and 30 dB SPL) increases systematically as the intensity of the added stimulus is increased but does not change within a condition as a function of session block. The former is predicted by the gain control model illustrated in Figure 1 but not by other models based on auditory attention. The latter suggests that practice reduces the variability along the decision axis but does not affect the operation of the gain control mechanism.

Although not shown, the fit of the model to the data of the six individuals is also good. Hence, the pattern of results expected from a gain control model: (1) characterizes individual as well as group data; (2) indicates that discriminability improves with practice; and (3) that characteristics of the soundscape (in this case, the highest intensity expected in the soundscape) control the amount of compression exerted on the input.

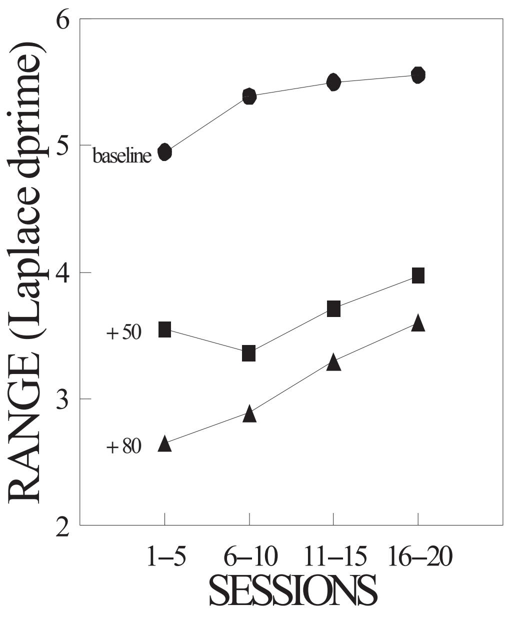

Figure 5. Laplace dprime range(distance between the means of distributions 1 and 4 in the EVLD Model) as a function of session block.

References

Gordon, M. S., & Schneider, B. A. (2007). Gain control in the auditory system. Absolute identification of intensity within and across two ears. *Perception & Psychophysics*, *69*, 232-240.

Parker, S., Murphy, D. R., & Schneider, B. A. (2002). Top-down gain control in the auditory system: Evidence from identification and discrimination studies. *Perception & Psychophysics*, *64*, 598-615.

Parker, S., & Schneider, B. (1994). Stimulus range effect: Evidence for top-down control of sensory intensity in audition. *Perception & Psychophysics*, *56*, 1-11.

Schneider, B. A. (2007). The shape of the underlying distributions in absolute identification experiments. In S. Mori, T. Miyaoka, & W. Wong (Eds.) *Fechner Day 2007: Proceedings of the 23rd Annual Meeting of the International Society for Psychophysics.* Tokyo, Japan: International Society for Psychophysics.

Schneider, B. A., & Parker, S. (1990). Does stimulus context affect loudness or only loudness judgments? *Perception & Psychophysics*, *48*, 409-418.

A CRITICAL-BAND-FILTER ANALYSIS OF JAPANESE SPEECH SENTENCES

Kazuo Ueda and Yoshitaka Nakajima
Perceptual Psychology Section, Kyushu University, Shiobaru, Minami-ku, Fukuoka 815-8540
Japan
ueda@design.kyushu-u.ac.jp nakajima@design.kyushu-u.ac.jp

Abstract

This investigation aims to seek common factors of speech sounds across different languages, which may be exploited widely in speech perception. Two-hundred sentences of Japanese, each uttered by 10 native speakers (5 females and 5 males), were analyzed through 20 bands of critical-band filters. Smoothed power fluctuations derived from the filters were submitted to factor analysis. The first three factors explained 33.6-34.8% of variance. These three factors, which could be related to four frequency bands, had appeared in the same way as in our previous analysis of British English speech. Intelligible noise-vocoded speech was obtained for both languages utilizing these frequency bands. The present analysis showed a common aspect of speech communication between Japanese and British English.

The present investigation focuses on frequency bands that adequately represent power fluctuation of critical-band-filtered Japanese sentences. This study is a continuation of our study on British English, presented at the last ISP meeting (Ueda & Nakajima, 2007).

The concept of critical bands (Fletcher, 1940) was proposed to model a frequency analysis function of our auditory system, more specifically the auditory periphery, in a simplified manner (Zwicker & Terhardt, 1980). The simplified model has been useful in estimating loudness of complex sounds, for example.

Critical-band filters have been also utilized in analysis of steady Dutch vowels. Plomp and his colleagues (Plomp, 1976) analyzed Dutch steady vowels with a bank of band-pass filters, which were practically equivalent to critical-band filters. They extracted principal components from the level fluctuation of the filter outputs. The first two principal components represented a vowel space very well, and the vowel configuration on a plane was well matched to the one obtained with the first two formant frequencies, the traditional way of analysis pioneered by Peterson and Barney (Peterson & Barney, 1952). However, Hillenbrand and his colleagues (Hillenbrand, Getty, Clark, & Wheeler, 1995; Hillenbrand & Nearey, 1999) clarified that spectral transition, together with static formant frequencies in steady portions of vowels, provides important cues to perceive vowels. The importance of spectral transition is also supported by other lines of investigation (Ladefoged & Broadbent, 1957; Verbrugge & Rakerd, 1986).

Studies on noise-vocoded speech suggest that the spectral transition can be rather coarse. The noise-vocoded speech (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995) is a degraded speech, in which the amplitude envelope of an original speech waveform is preserved but the fine structure is replaced with noise. Usually amplitude envelopes are extracted from band-pass filtered outputs of the original speech sound. Thus, coarse mapping of spectral transition into several frequency bands is included in a process of synthesizing a noise-vocoded speech. Generally, intelligibility of noise-vocoded speech depends on the number of frequency bands: as the number increases, the speech becomes more intelligible. With extensive training--without feedback, however--noise-vocoded speech synthesized with four successive band-pass filters (that means the whole frequency range of speech is divided into four frequency bands) can be fairly intelligible, i.e., can yield about 85% of word accuracy (Dorman, Loizou, & Rainey, 1997; Shannon et al., 1995; Smith, Delgutte, & Oxenham, 2002).

Thus, our auditory system seems to be able to utilize a small number of frequency channels, each of which has a much broader bandwidth than a critical bandwidth, when perceiving running speech. Our previous investigation on British English (Ueda & Nakajima, 2007) revealed that a whole frequency range of speech could be divided into four frequency bands, according to factor analyses of power fluctuations derived through critical-band filters. The purpose of the present investigation is to analyze Japanese speech sentences in the same manner, and to compare the results with those of British English.

Method

Speech Samples

Two-hundred speech sentences, each uttered by 10 Japanese speakers (5 females and 5 males), were used. Those materials were included in a commercial speech database (NTT-AT, "Multilingual Speech Database, 2002") with 16-kHz sampling and 16-bit quantization. One of the authors edited the materials to eliminate blanks and noises, by using a computer program developed by the authors. Additionally, 200 speech sentences of British English, each uttered by 4 English speakers (2 females and 2 males), were analyzed. These were included in another speech database (ATR, "The ATR British English Speech Database") with 12-kHz sampling and 12-bit quantization.

Analyses

Figure 1 shows a block diagram of the analyses. Two banks of critical-band filters, A and B, were constructed. The total pass-band of the filter bank A ranged 20-6400 Hz, and the corresponding center frequencies ranged 50-5800 Hz. Cutoff frequencies of each filter were determined according to Zwicker and Terhardt (Zwicker & Terhardt, 1980), except the lowest cutoff frequency. Cutoff frequencies of the filter bank B, covering 50-7000 Hz with center frequencies of 100-6400 Hz, were halfway shifted from those of bank A, in order to check an effect of frequency settings. Each filter output was squared, smoothed with a 10-ms Gaussian window, and sampled at every 1 ms. The data were pooled over each individual, the females, the males, and all the speakers, and submitted to factor analyses.

Results

Figure 2 shows the results for Japanese. The cumulative contributions were 33.61% and 34.78% (filter banks A and B, respectively) for all the speakers, 32.29% and 33.58% for the females, and 39.98% and 40.45% for the males. Figure 3 shows the results for British English. The cumulative contributions were 41.80% and 40.32% for the ATR database, which included 4 speakers, and 35.57% and 36.55% for the NTT-AT database, which included 10 speakers (Ueda & Nakajima, 2007). These results were remarkably similar to each other. There were only marginal differences between the results obtained with filter bank A and filter bank B.

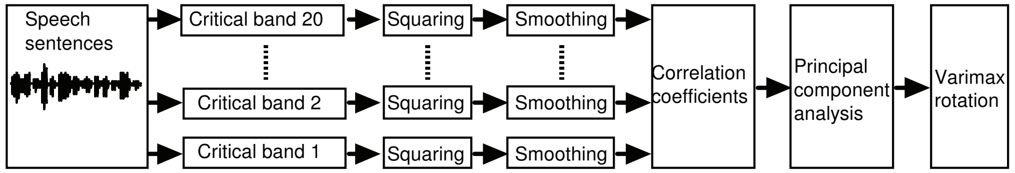

Fig. 1. A block diagram of the analyses.

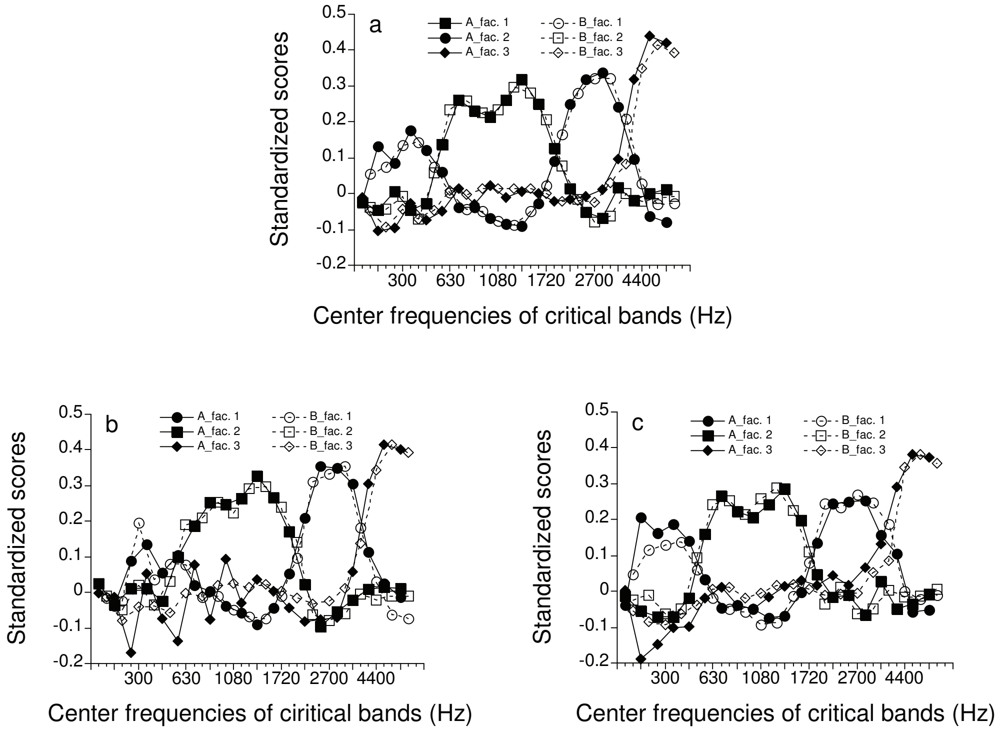

Fig. 2. Standardized scores of the first three factors obtained from pooled data of (a) all the Japanese speakers, (b) the females, and (c) the males. The solid lines and the dashed lines represent the results obtained with critical-band filter banks A and B, respectively.

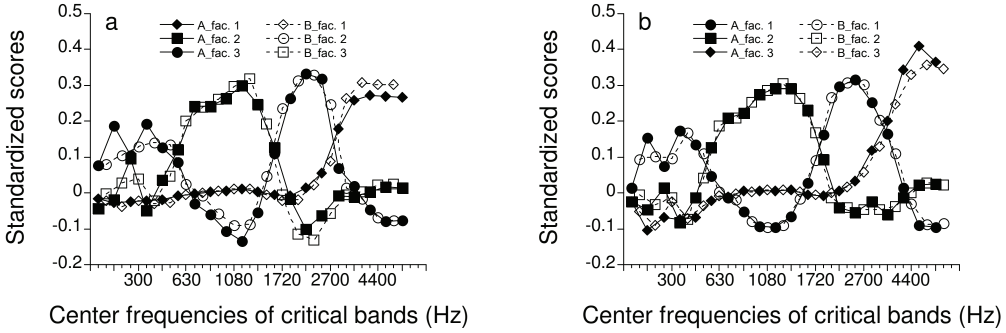

Fig. 3. Standardized scores of the first three factors of British English speakers: (a) ATR database and (b) NTT-AT database (Ueda & Nakajima, 2007).

Discussion

Two points were clarified: (1) quite similar patterns of results were obtained over two distinctly different languages, i.e., Japanese and British English, and (2) the analyses of British English were highly replicable over two independent speech databases (NTT-AT and ATR). To assess boundaries of frequency bands, we took the crossover frequencies of the curves in the figures. The boundaries were 510, 1880, and 2700 Hz in Japanese, and 550, 1800, 3300 Hz in British English of the NTT-AT database. Informal listening tests by the authors showed that either set of boundaries could yield intelligible noise-vocoded speech in both languages. Therefore, these four frequency bands should represent fundamental processing units along frequency axis related to speech perception.

Acknowledgements

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 and by a Grant-in-Aid for the 21st Century COE program.

References

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.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97*, 3099-3111.

Hillenbrand, J. M., & Nearey, T. M. (1999). Identification of resynthesized /hVd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America, 105*, 3509-3523.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*, 98-104.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*, 175-184.

Plomp, R. (1976). *Aspects of Tone Sensation: A Psychophysical Study*. London: Academic Press.

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

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

Ueda, K., & Nakajima, Y. (2007). Critical-band filter analysis of speech sentences, *The 23rd Annual Meeting of the International Society for Psychophysics* (pp. 503-508). Tokyo.

Verbrugge, R. R., & Rakerd, B. (1986). Evidence of talker-independent information for vowels. *Language and Speech, 29*, 39-57.

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

SPEECH- SYNCHRONIZED VISUAL CUES RELEASE SPEECH FROM INFORMATIONAL MASKING

Mengyuan Wang, Jingyu Li, Liang Zhang, Yanhong Wu, Xihong Wu, and Liang Li
Department of Psychology, Peking University, Beijing, China, 100871
motoluto@163.com

Abstract

Visual speech information, such as lipreading cues, assists listeners to segregate a target voice from competing voices. It is not clear whether a simple visual cue, such as the light flash that is synchronous to the onset of each syllable in target speech, is sufficient to release target speech from noise or speech masking. In this study, when target speech was of a constant rate, the speech-synchronized light flash had no unmasking effects. However, when the rate of target speech was artificially manipulated, the speech-synchronized light flash improved speech recognition when the two-talker speech masker but not the speech-spectrum noise masker was co-presented. Thus, under certain conditions, speech-synchronized visual cues can play a role in helping listeners attend to the target voice and follow the stream of target speech, leading to a release of target speech from informational masking.

People often participate in conversations in noisy environments with noise sounds and person talking. Under such adverse conditions, listeners with normal hearing can use some perceptual cues to segregate target speech from the noise background. For example, viewing a speaker’s articulatory movements (e.g., lipreading) substantially improves a listener’s recognition of the speaker’s speech especially under noisy conditions. Helfer and Freyman (2005) have recently reported that the effect of lipreading on speech recognition is masker-type dependent. Lipreading can only release speech from speech masking but not noise masking, suggesting that visual cues help listeners overcome informational masking but not energetic masking.

However, lipreading information is very complicated. This study investigated whether a single-dimensional signal in lipreading, the speech-synchronized light flash (which temporally matched the onset of each syllable in a target speech sentence) is sufficient to unmask speech.

Participants

Thirty-six young university students participated in this study, twelve in Experiment 1 and twenty-four in Experiment 2 (twelve in each part of Experiment 2). They had normal and symmetrical hearing (no more than 15 dB difference between the two ears, pure-tone hearing thresholds < 25 dB HL between 0.125 and 8 kHz). Their first language was Mandarin Chinese.

Apparatus

The participant was seated at the center of an anechoic chamber (Beijing CA Acoustics). All acoustic and visual signals were digitized using the 24-bit Creative Sound Blaster PCI128 and audio editing software (Cooledit Pro 2.0). The acoustic analog outputs were delivered from a loudspeaker (Dynaudio Acoustics, BM6 A) 200 cm in front of the participant. The flash was delivered from a light-emitting diode (LED) at the center of the loudspeaker.