CRITICAL-BAND FILTER ANALYSIS OF SPEECH SENTENCES

Kazuo Ueda and Yoshitaka Nakajima
Perceptual Psychology Unit, 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

Two-hundred English sentences, each spoken by 10 native speakers (5 females and 5 males) of British English, were analyzed with a bank of 20 critical-band filters. Principal component analysis was applied to the power fluctuation of filter outputs. The first three components accounted for 35.3% of the total variance. A varimax rotated solution of the components was obtained. The first component was interpreted as a "sonority filter" that is closely related to the first and the second formants of vowels. The second component seemed to be related with nasalization, and the third component with fricatives, affricates, and stops. The existence of common frequency channels in perceiving normal speech and noise-vocoded speech were suggested.

The concept of critical band is one of the successful simplifications of an aspect of our auditory functions. The concept is originally proposed by Fletcher (1940). He proposed that critical band is dependent on the auditory periphery, particularly on the excitation patterns of the basilar membrane in the cochlea. The concept had been further simplified to be modeled as a bank of band-pass filters that do not overlap each other (e.g., Fastl & Zwicker, 2007). This model is well matched with the essential parts of the experimental facts of auditory masking and loudness perception.

The concept of critical band has been also applied to the research on speech perception. Plomp and his colleagues (Klein, Plomp, & Pols, 1970; Plomp, Pols, & van de Geer, 1967; Pols, Tromp, & Plomp, 1973; Pols, van der Kamp, & Plomp, 1969) analyzed steady portions of Dutch vowels with a bank of band-pass filters that are comparable to critical bands. They applied principal-components analysis to the level fluctuations of the filter outputs and found that a small number of factors reasonably explained the variance, and that the obtained vowel configuration could be well matched to the configuration on a F1-F2 (the first and the second formants) plane.

Both temporal information and spectral information are vitally important in perceiving speech sentences. Noise-vocoded speech of only four frequency bands, i.e., four-bands of noise each modulated by an amplitude envelope of a filtered speech with a corresponding pass-band, can be well recognized despite its severe degradation in spectrum (Apoux & Bacon, 2004; Riquimaroux, 2006; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995; Shannon, Zeng, & Wygonski, 1998). The fact that a single-band of noise-vocoded speech is almost unrecognizable (Shannon et al., 1995) suggests that appropriate spectral information is necessary to recognize speech, especially vowels and speech sentences (Shannon et al., 1998).

Nakajima and Ueda (2006) proposed an index for detecting syllabic boundaries (syllabic boundary index) that is based on power ratios among the outputs of three time-frequency windows shown in Figure 1. They showed a possibility to detect syllabic boundaries based on the index. They also demonstrated that consonant enhancement can be

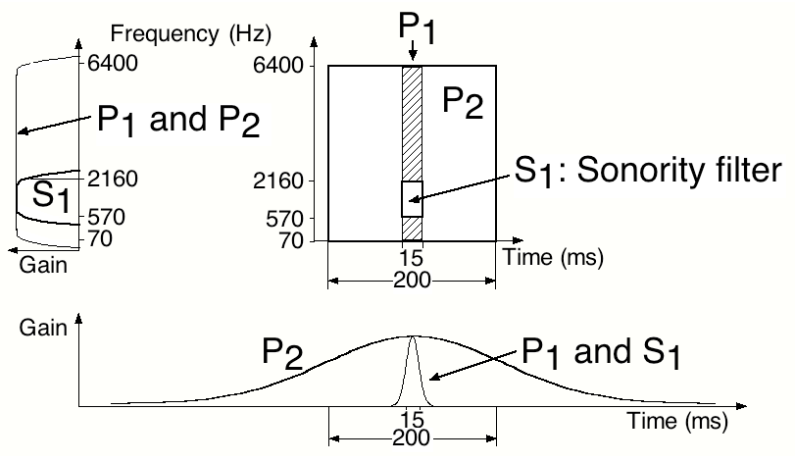

Fig. 1. Time-frequency windows for detecting syllabic boundaries.

achieved by utilizing the power ratios for controlling the degree of amplification around consonants. One of the time-frequency windows has a pass-band of 570-2160 Hz, which roughly corresponds to the frequency range of the first and the second formant frequencies of vowels. They called that window *sonority filter*, which is named after the founder of modern linguistics, Ferdinand de Saussure, who investigated the relationship between *sonority* and syllabic boundaries (de Saussure, 1916/1959).

Those researches on noise-vocoded speech and syllabic boundary index seem to suggest that there may be some fixed frequency channels that are vitally important in perception of running speech and in detecting syllabic boundaries. Moreover, those small numbers of channels can be represented as some groups of critical bands. To our knowledge, however, no such research has been undertaken. The purpose of our investigation is to find principal components in power fluctuations of critical-band-filtered speech sentences.

Method

Two-hundred English sentences each uttered by 10 native speakers of British-English (5 females and 5 males) were taken from a speech database (NTT-AT, 2002). The sentences read were based on those appeared in English journals and magazines. The speech sound was

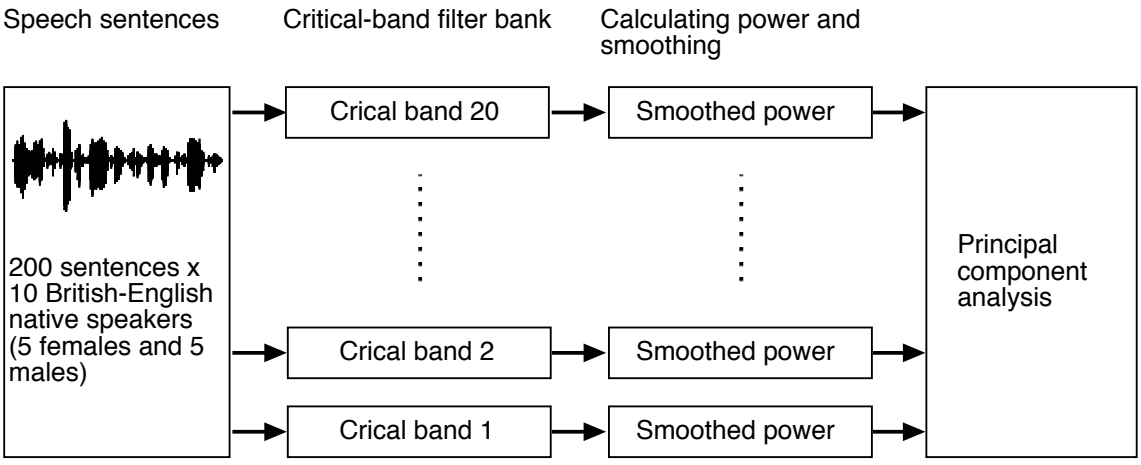

Fig. 2. Block-diagram of the analysis.

recorded with a 16-kHz sampling frequency and a 16-bit quantization.

Each recording contains relatively long blanks before and after a speech sentence. Sometimes noise associated with breathing and mouth opening was also included. One of the authors edited all the speech files to eliminate long (more than about 10 ms) blanks and noises before and after a speech sentence, using a hand-made editing program. The program enabled semi-automatic elimination of blanks and noises, and to compare waveforms both from an original and an edited file on a computer screen. When the operator detected a portion that was suspected to be a failure of editing, he examined the waveform again, listened to a suspected portion if he thought it was necessary, adjusted the editing parameters, and edited the file again. When the whole editing process was once over, the operator inspected the entire waveforms comparing original one and edited one, to ensure proper editing.

A block-diagram of the analysis is shown in Figure 2. A bank of 20 critical-band filters was constructed according to the pass-band and the center frequencies described by Fastl and Zwicker (2007). A slope of each filter corresponded to about -300 dB/oct. The smoothed power fluctuation was calculated according to the following three steps: each filter output was squared, taken the moving average with a 1-ms rectangular window, and taken the moving average with a 10-ms Gaussian window. The results of calculation were sampled at every 1 ms, combined for each speaker and for each gender group, and finally integrated into a large file of whole speakers. These data are submitted to principal component analysis by

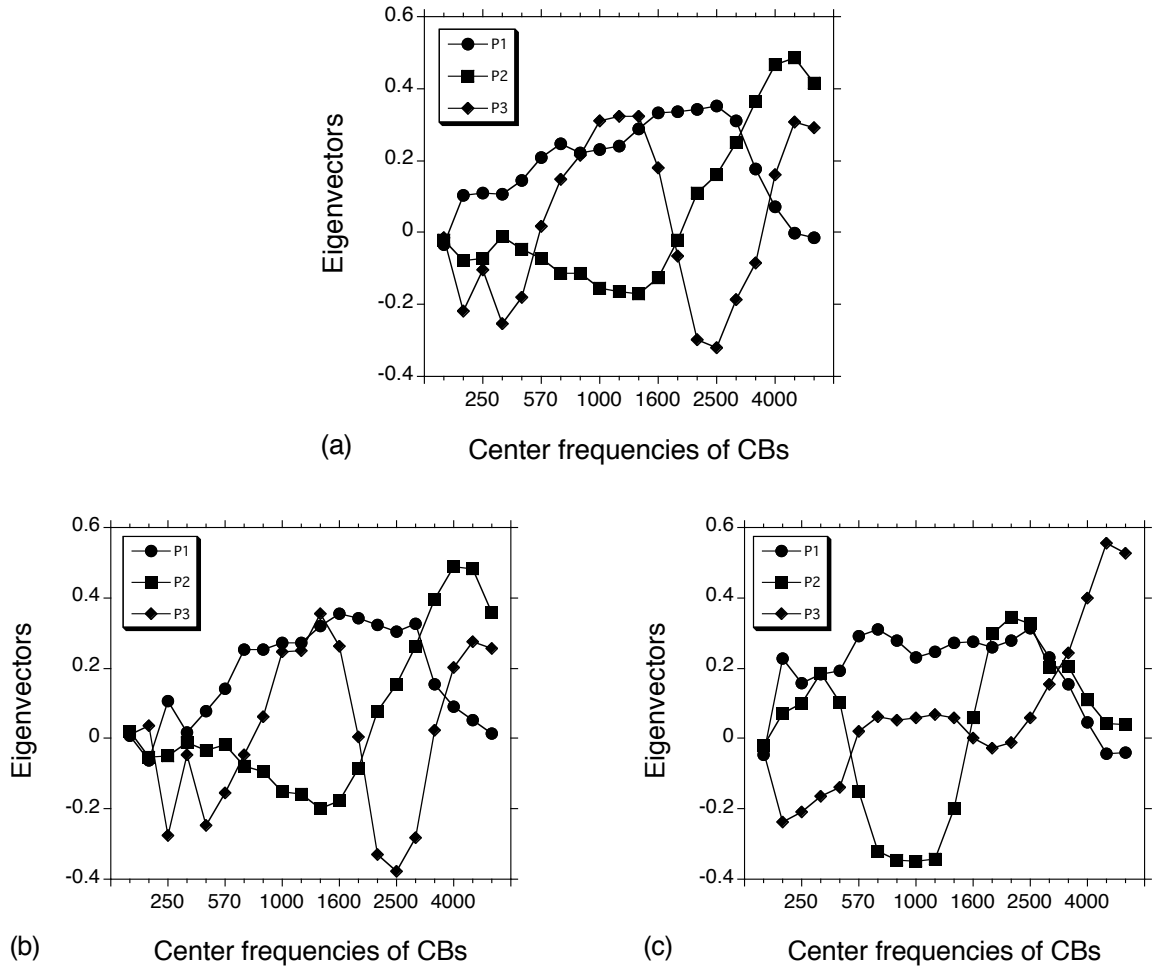

Fig. 3. Eigenvectors of the first three principal components (P1-P3): (a) pooled data, (b) female speakers, and (c) male speakers.

using JMP 6.0.3 (SAS Inst. Japan).

Results and Discussion

Cumulative contribution of the first three principal components was 35.3% of the total variance. Figure 3 shows eigenvectors of the first three principal components, and Figure 4 shows standardized scores of the same components obtained with a varimax rotation.

It is interesting to see the curve of the eigenvector of the first principal component (Fig. 3) corresponds well to the pass-band of a conventional telephone (300-3200 Hz), if one regards the curve as a frequency response of a filter. Varimax rotated scores in Figure 4 show better consistency between speaker gender groups than eigenvectors in Figure 3; although the order of the second and the third components are substituted for the female speakers and for the male speakers, the shapes of the curves in the figures are very consistent, especially when one focuses on the crossover frequencies of those curves: 570, 1850, and 3150 Hz in Figure 4a, 700, 1720, and 3150 Hz in Figure 4b, and 510, 1480, and 3150 Hz in Figure 4c.

Those standardized scores can be regarded as weights for critical-band filter outputs, and thus an example of the weighted outputs based on the scores in Figure 4a were realized as the spectrograms shown in Figure 5. The first principal component after a varimax

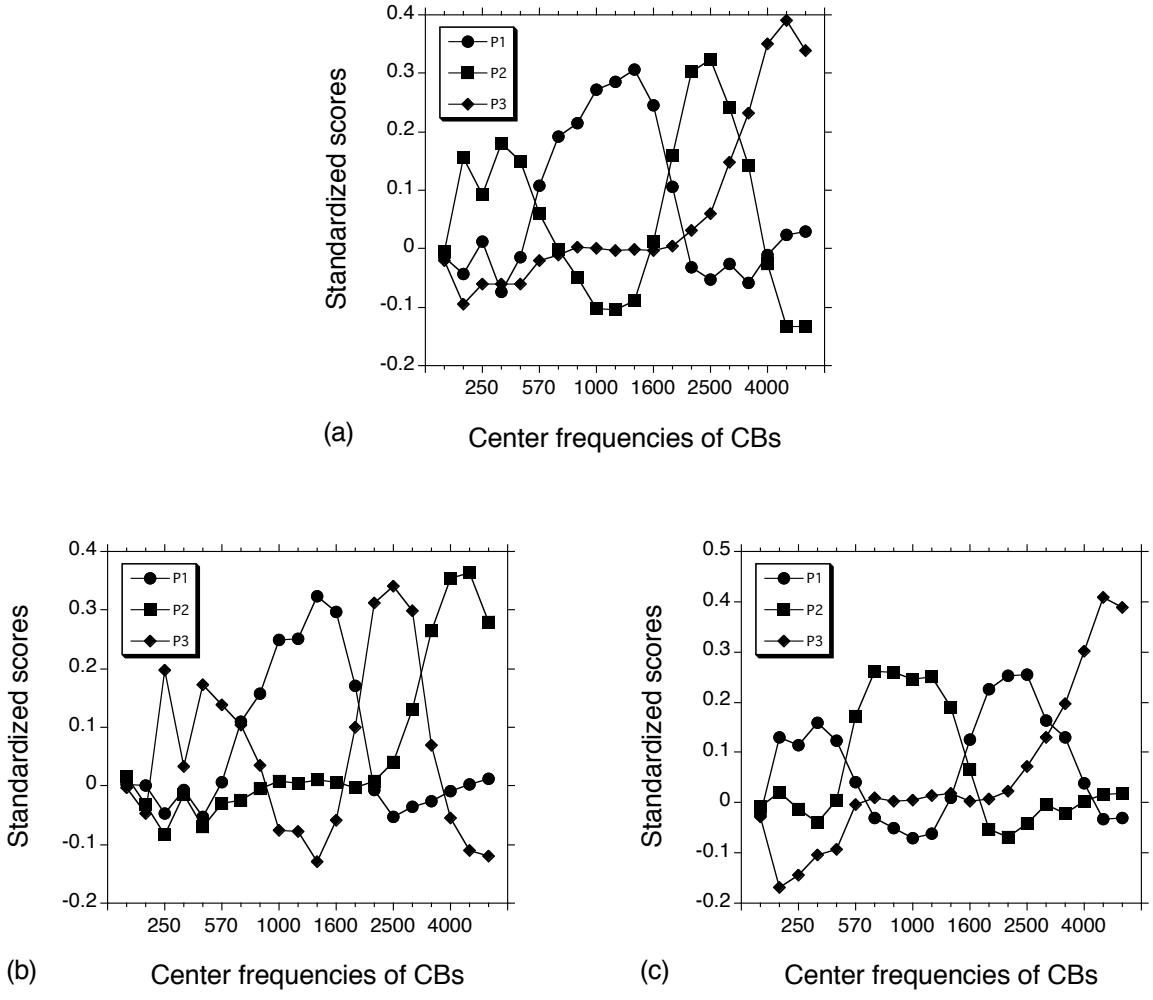

Fig. 4. Standardized scores of the varimax rotated principal components (P1-P3): (a) pooled data, (b) female speakers, and (c) male speakers.

rotation is well corresponding to a sonority filter. The second principal component seems to represent nasalization. The third principal component is supposed to represent consonants such as fricatives, affricates, and stops, like /f, v, s, z, t, tʃ, d, dʒ, p, k, g/ etc. These three components divide the whole frequency range of speech into four channels. A noise-vocoded speech also requires four bands of frequency channels to attain maximum sentence recognition: the filter cutoff frequencies were set at 800, 1500, and 2500 Hz in Shannon and his colleagues' case (Shannon et al., 1995; Shannon et al., 1998), and 600, 1500, and 2100 Hz in Riquimaroux and his colleagues' case (Riquimaroux, 2006). Although these cutoff frequencies are somewhat different from the crossover frequencies in the present results, it is worth noting that the number of frequency channels coincides with each other: it is possible that these frequency channels are used in recognizing both normal and degraded speech. Moreover, perceptual grouping may occur with the lowest and the third frequency channels, because these two channels roughly correspond to the nasality component of the present

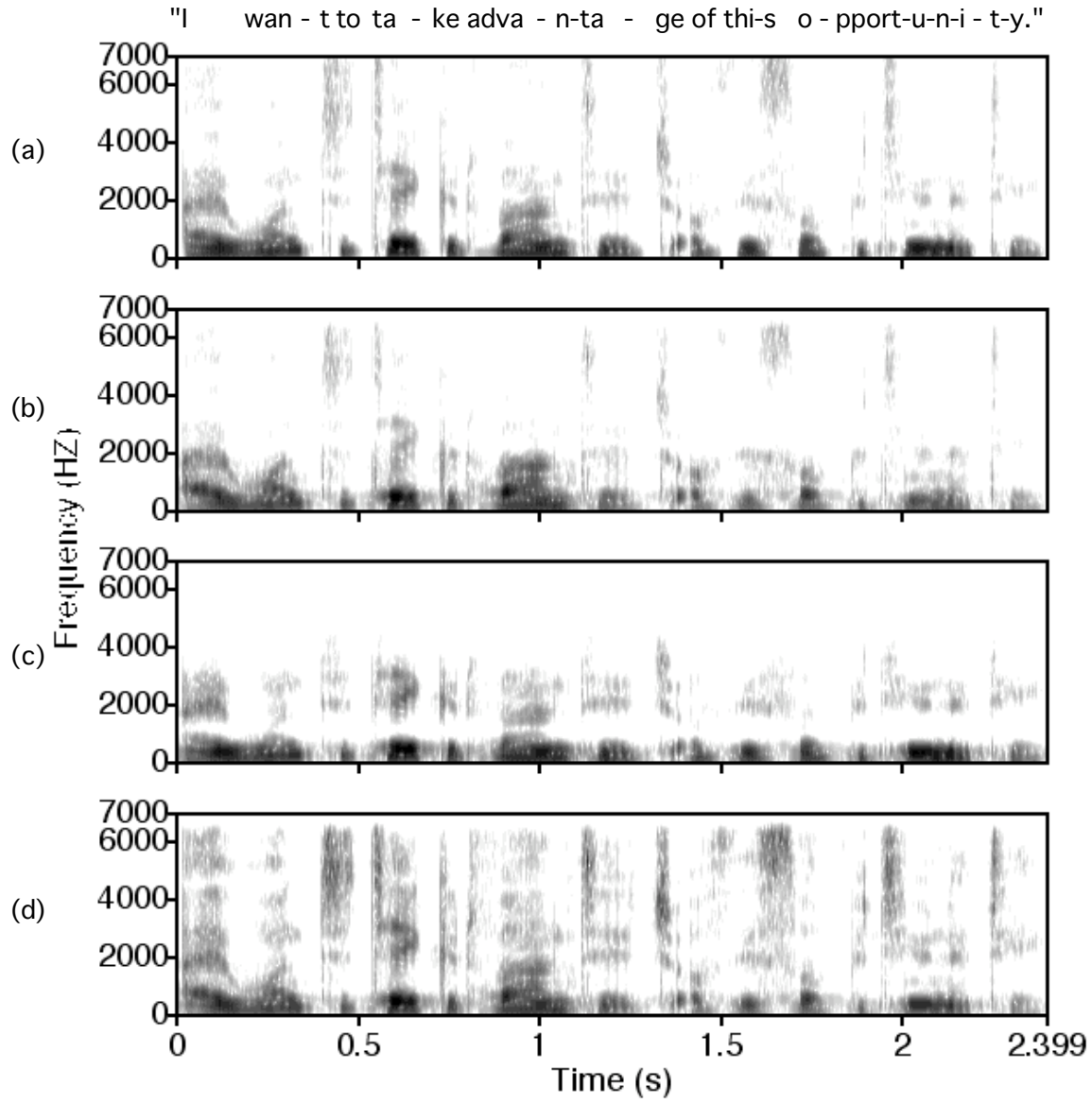

Fig. 5. An example of weighted spectrograms according to the standardized scores: (a) the original speech sentence spoken by a female speaker, and (b)-(d) weighted spectrograms with the first, the second, and the third principal components, respectively.

results.

Acknowledgements

This research was supported by Grants-in-Aid for Scientific Research Nos. 14101001 and 19103003 from the Japan Society for the Promotion of Science and by a Grant-in-Aid for the 21st Century COE program.

References

Apoux, F., & Bacon, S. P. (2004). Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise. *Journal of the Acoustical Society of America, 116*, 1671-1680.

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

Fastl, H., & Zwicker, E. (2007). *Psychoacoustics: Facts and Models* (Third ed.). Berlin: Springer.

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

Klein, W., Plomp, R., & Pols, L. C. W. (1970). Vowel spectra, vowel spaces, and vowel identification. *Journal of the Acoustical Society of America, 48*, 999-1009.

Nakajima, Y., & Ueda, K. (2006). Auditory events in language and music. *Journal of the Acoustical Society of America, 120*, 3166.

NTT-AT. (2002). Multi-lingual speech database 2002. Tokyo: NTT-AT.

Plomp, R., Pols, L. C. W., & van de Geer, J. P. (1967). Dimensional analysis of vowel spectra. *Journal of the Acoustical Society of America, 41*, 707-712.

Pols, L. C. W., Tromp, H. R. C., & Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *Journal of the Acoustical Society of America, 53*, 1093-1101.

Pols, L. C. W., van der Kamp, L. J. T., & Plomp, R. (1969). Perceptual and physical space of vowel sounds. *Journal of the Acoustical Society of America, 46*, 458-467.

Riquimaroux, H. (2006). Perception of noise-vocoded speech sounds: Sentences, words, accents and melodies. *Acoustical Science & Technology, 27*, 325-331.

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

Shannon, R. V., Zeng, F.-G., & Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *Journal of the Acoustical Society of America, 104*, 2467-2476.