ACOUSTIC CORRELATE OF PHONOLOGICAL SONORITY IN BRITISH ENGLISH

Yoshitaka Nakajima*, Kazuo Ueda*, Shota Fujimaru*, Hirotoshi Motomura*, Yuki Ohsaka*
*Kyushu University, Fukuoka, Japan
nakajima@design.kyushu-u.ac.jp

Abstract

Sonority or aperture proposed in linguistic literature can be considered a kind of subjective measure specific to speech perception. Vowels have high sonorities corresponding to a linguistic fact that they can be nuclei of syllables, and fricatives and stops have low sonorities. In order to understand how sonority is perceived, we attempted to find an acoustic dimension on which we could construct a psychophysical scale of sonority. We applied the multivariate analysis method as in Ueda et al. (2010, Fechner Day, Padua) to spoken sentences in British English collected in a commercial database, in which phonemes were segregated and labeled. The speech signals went through a bank of critical-band filters, and the output power fluctuations were subjected to factor analysis. The three factors as in our previous study appeared. The analyzed phonemes were classified into three categories, i.e., vowels, sonorant consonants, and obstruents. These categories were represented well in the Cartesian space whose coordinates were the factor scores of the above factors. One of the factors located around 1000 Hz was highly correlated with sonority or aperture.

The concept of the syllable is important in accounting for why the temporal order of phonemes is often guided by a set of rules or constraints (e.g., Spencer, 1996; Prince & Smolensky, 2004). This study was originally an attempt to find a psychoacoustic basis to understand how syllables were formed. We were interested in whether the three-factor representation of speech sounds as in Ueda et al. (2010) could be related to their phonological categories. Vowels are often represented in a formant map, and this helps a lot to understand how vowels are articulated and perceived. We took a further step: We took up spoken sentences in British English, and tried to draw a map of vowels, including diphthongs, and consonants put together on a purely acoustic basis. Most consonants and all diphthongs are characterized by systematic spectral changes in time, and these changes are often important for speech perception. However, we did not take up such spectral changes in the present stage because they seemed too complicated to be reflected in a simple map. We thus took up the following strategy. We specified representative factor scores for each observed phoneme as a first step. We then observed the results to judge whether they reflected essential aspects of syllable formation. If so, then we could take a next step, and to analyze spectral changes within phonemes would be one possible alternative. If not so, then the present analysis would not be promising, and another way should be looked for. Now, the first step resulted in a configuration of phonemes which seemed worth reporting as it was. Thus, we report how the configuration was obtained and interpreted.

We looked for a database of speech sounds in which most phonemes of a certain language appeared, and were segregated and labeled. Fortunately, a commercial database of British English [The ATR British English speech database (Campbell, 1993)] was available to study this issue. This database was developed for speech science research. British English would give us a secure starting point, because its phonology has been described thoroughly in the literature (e.g., Harris, 1994; Spencer, 1996). About 25,000 samples of phonemes uttered by three speakers were available; the sheer amount of data was a great advantage of utilizing this database.

Analysis

We analyzed spoken sentences in British English to extract the three factors to explain part of the power fluctuations in twenty neighboring critical bands. We determined factor scores of each labeled phoneme at the temporal middle of its labeled duration.

Method

Speech samples The database comprises two-hundred English sentences read aloud by three native speakers of British English: two females and one male (There was another male speaker, but there was a technical problem in the labeling data of this speaker, which could not be fixed by the company that released the database). The speech signals were recorded with 12-kHz sampling and 12-bit quantization. Labeling was performed manually; all the phoneme labels were linked to specific periods of speech signals that sounded approximately as indicated if played separately.

The labeling data sometimes did not completely agree with the phonemes indicated in dictionaries expressing subtle differences of the sounds, but such differences were not within our present interest, which was to connect the acoustic and the phonological features of the speech signals. The labeling data thus had to be modified in some cases. We preserved the original labeling data of the database as far as possible, but incoherent cases were omitted. There were

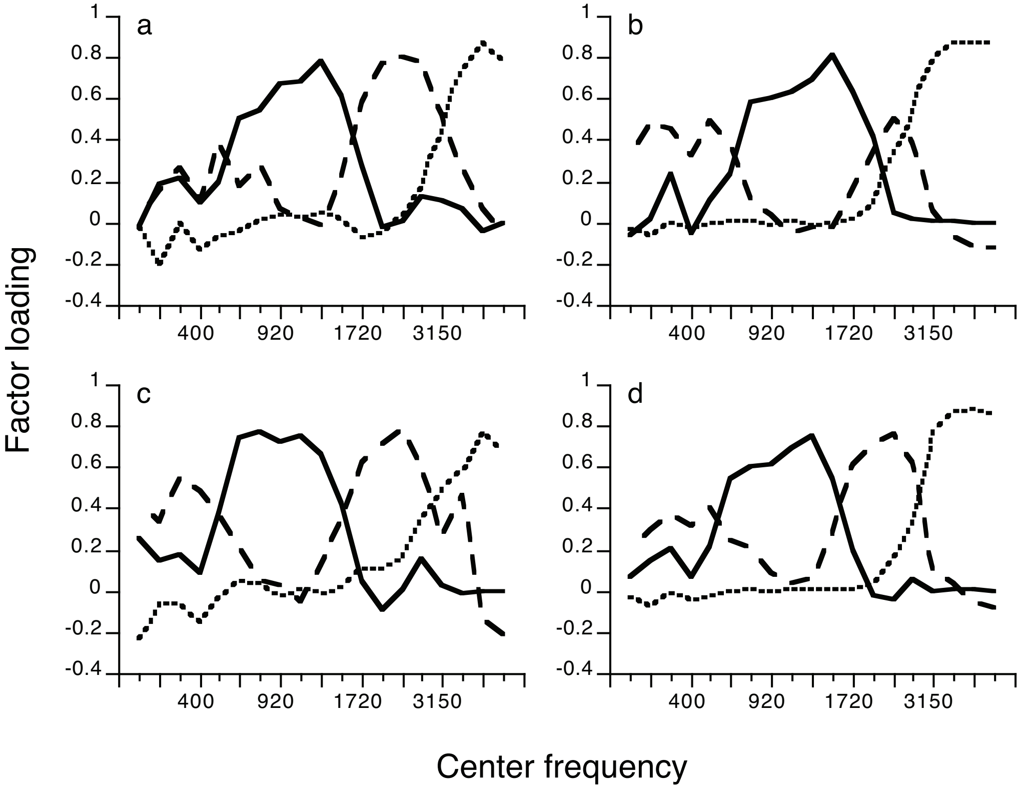

Figure 1 Obtained factors. Panels (a) and (b) show the results for two female speakers, respectively, and panel (c) shows the results for a male speaker. The results for all the speakers are combined in panel (d).

cases, although few, in which voiced and unvoiced phonemes were exchanged in the labeling data; these cases were not omitted, and the phoneme labels in the database were utilized. Some phonemes were separated into more than one period in the database. One representative period was chosen in such a case; a closure part was always omitted from further analyses, even when it was the only part representing a phoneme, because closure parts had almost no sound energy to differentiate them from each other. When a vowel was separated into a main part and a transient nasalized part, the main part was chosen. If a stop consonant was still separated into more than one period, the period in which a stronger aspiration was indicated in the label was chosen. Out of the 31 663 labeled periods in the database, 6 754 periods were omitted in the above screening.

Procedure All the speech signals were analyzed as in Ueda et al. (2010). Power fluctuations were derived from 19 critical-band filters (Zwicker & Terhardt, 1980). The lowest critical band was 50-150 Hz; spectral components below 50 Hz were neglected because no substantial parts of speech signals were to be expected in this range. The 19th, i.e., the highest, critical band was 4800-5800 Hz. The derived power fluctuations were submitted to principal component analyses, and factors were determined by varimax rotation.

Each labeled period of the database was then connected to a set of factor scores. The outputs of the critical-band filters were utilized in this calculation; the output powers at the temporal middle of the period were converted into factor scores by applying the factor loadings obtained in the above analysis.

Thus, all the labeled phonemes as screened above were represented in a Cartesian space of the factor scores, i.e., a factor space. Finally, each English phoneme was represented by a single point in the space; the sets of factor scores were averaged for each English phoneme. A configuration of the English phonemes was thus obtained by a purely acoustic analysis; exactly speaking, the labeling data in the speech database were the only linguistic information utilized in the present analysis.

Results

The varimax rotation led to three factors to be related to four frequency ranges as in Figure 1. The speech signals of the three speakers were combined and submitted to a single analysis

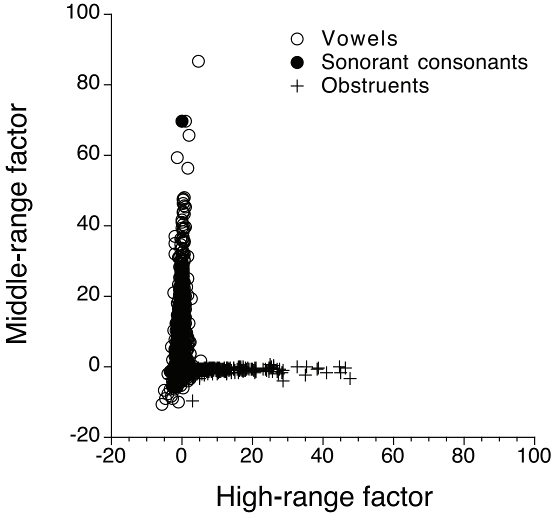

Figure 2 An L-shaped distribution of labeled phonemes on a plane of the high-range and the middle-range factor.

(Figure 1d), and the speech signals of individual speakers were also analyzed separately (Figures 1a-c). The cumulative contributions of the three factors were 41-45% in all these analyses. The following three factors appeared: the *high-range factor* appeared above 3200 Hz, the *middle-range factor* around 1100 Hz, and the *exterior factor* in two separate frequency ranges around 300 and 2200 Hz. The results were consistent with our previous results (Ueda et al., 2010).

The labeled phonemes were represented in a three-dimensional factor space corresponding to the above factors. Their configuration was characteristically L-shaped on the plane of the high-range and the middle-range factor (Figure 2). The high-range factor seemed to be related only to obstruents (fricatives, affricates, and stops), and the middle-range factor only to vowels and sonorant consonants (glides, liquids, and nasals). Although vowels and sonorant consonants shared a substantial area, their distributions differed; vowels could take higher scores of the middle range factor, and the distribution of sonorant consonants were often limited to an area that was shared by vowels but related to smaller values of the middle-range factor scores.

The high-range factor and the middle-range factor thus could be expressed as a single dimension by observing the origin of the space from a viewpoint in the space in which both factors took positive values (Figure 3). Vowels and sonorant consonants were separated clearly from obstruents.

Each English phoneme was represented by the average position (center of gravity) of all the labeled samples in this space (Figure 4). In this presentation, the phonemes were clearly separated into three categories: vowels, sonorant consonants, and obstruents.

Discussion

The obtained configuration of the phonemes seems to have a close relationship with English syllable formation. Most English vowels can be nuclei of stressed syllables, and they are distributed in the upper right part of the configuration. On the contrary, obstruents can never be syllable nuclei, and they are distributed in the lower left part. Sonorant consonants, which are considered more vowel-like than obstruents, are located near the middle of the configuration; note that /l/ and /n/ can be nuclei of unstressed syllables.

When the factor scores of the high-range and the middle-range factor are near or below

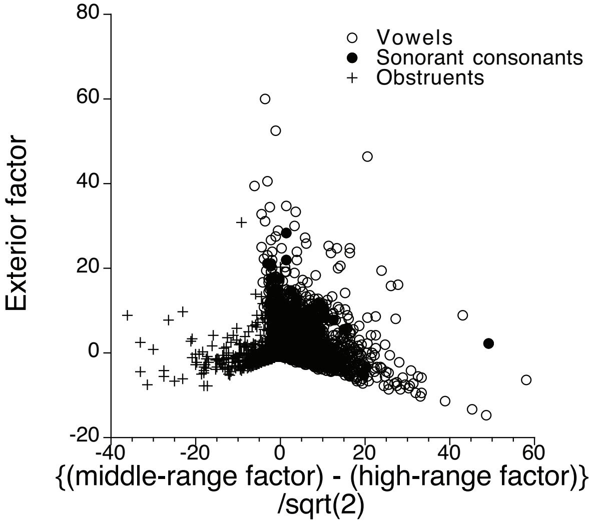

Figure 3 Distribution of labeled phonemes in the factor space.

zero, as is observed near and above the origin in Figure 4a, the factor score of the exterior factor can change in a wider range; vowels are located in the upper part, and obstruents in the lower part.

The nature of phonemes related to the gradual change through vowels, sonorant consonants, and obstruents is called *sonority*, or *aperture*, in linguistics. Typically, Spencer (1996) proposed the following sonority scale, in which larger numbers indicate higher sonority: 6) vowels, 5) glides, 4) liquids, 3) nasals, 2) fricatives/affricates, 1) plosives (stops). The English

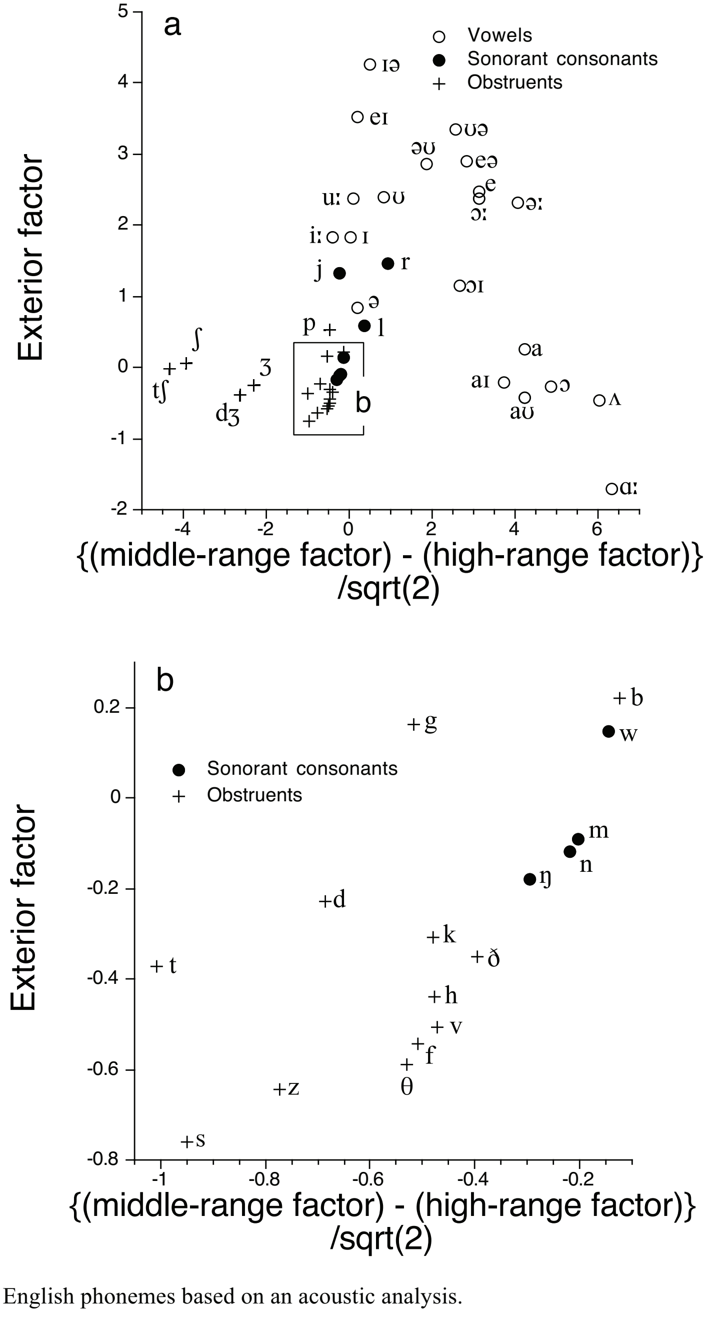

Figure 4 A map of English phonemes based on an acoustic analysis.

phonemes represented in the above factor space can be related to this sonority scale. The Spearman's rank correlation coefficient between the middle-range factor score and the sonority scale value was as high as 0.83 (N=44; $p<0.05$). The other two factors also showed significant rank correlations: -0.45 for the high-range factor ($p<0.05$) and 0.54 for the exterior factor ($p<0.05$). For a classic example, de Saussure (1959) proposed a similar scale of aperture (in an appendix, which is rarely quoted), and the aperture value showed a significant rank correlation (0.83; N=31) with the middle factor score. Considering the factor loadings related to the center frequencies of the critical-band filters (Figure 1), sonority is very likely to be a subjective dimension corresponding, psychophysically, to the acoustic power below 2500 Hz; the acoustic power around 1000 Hz seems especially important, creating the impression of stressed syllables. The acoustic power above 4000 Hz may work to suppress sonority, making the impression of syllable boundaries clearer.

General Discussion

Sonority, which has been a purely linguistic concept created mainly to understand phonological rules or constraints especially in syllable formation, is now a concept that seems useful in computational and psychophysical approaches to auditory perception. The middle-range factor located around 1000 Hz can be a first approximation of a physical property to be put against sonority in psychophysical studies. It is also important that frequency components above 4000 Hz may be playing important roles in speech perception. Such components are neglected in the present telephone communication, and we are now developing a system in which the information carried by the high frequency components can be transmitted via a lower frequency channel.

Acknowledgments. Ger Remijn, Valter Ciocca, Mark Elliott, and Peter Howell gave us valuable suggestions. This research was supported by grants to YN and KU from the Japan Society for the Promotion of Science (19103003, 20653054) and from Kyushu University (an A type project; The Kyushu University Project for Interdisciplinary Research of Perception and Cognition).

References

Campbell, N. (1993). The ATR British English speech database. *Technical Report*, ATR Interpreting Telephony Research Labs.

de Saussure, F. (1959). *Course in General Linguistics*. (Baskin, W., Trans.). McGraw-Hill, New York (Original work published 1916).

Harris, J. (1994). *English Sound Structure*. Blackwell, Oxford.

Prince, A. & Smolensky, P. (2004). *Optimality Theory: Constraint Interaction in Generative Grammar.* Blackwell, Malden.

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

Ueda, K., Nakajima, Y., & Satsukawa, Y. (2010). Effects of frequency-band elimination on syllable identification of Japanese noise-vocoded speech: Analysis of confusion matrices. *Fechner Day 2010*, 39-44.

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 Amarica, 68*, 1523-1525.