How sonority appears in speech analyses

Yoshitaka Nakajima1,*, Kazuo Ueda1, Gerard B. Remijn1, Yuko Yamashita2 and Takuya Kishida3

1Department of Human Science/Research Center for Applied Perceptual Science, Kyushu University, 4–9–1 Shiobaru, Minami-ku, Fukuoka, 815–8540 Japan
2School of Arts and Sciences, Shibaura Institute of Technology, 307 Fukasaku, Minuma-ku, Saitama, 337–8570 Japan
3Graduate School of Design, Kyushu University, 4–9–1 Shiobaru, Minami-ku, Fukuoka, 815–8540 Japan

Keywords: Speech, Spectral changes over time, Multivariate analysis, Phonology, Intelligibility

PACS number: 43.66.Ba, 43.66.Mk, 43.70.Kv, 43.71.Es, 43.71.Ft, 43.71.Gv, 43.71.Hw, 43.71.Sy, 43.72.Ar, 43.72.Ja [doi:10.1250/ast.39.179]

1. BACKGROUND

Speech signals are almost perfectly intelligible even if they are reduced to power fluctuations in narrow frequency bands simulating functions of the auditory periphery, typically critical-band filters (Zwicker & Terhardt, 1980) [1]. If we take up an audio frequency range up to 7,000 Hz, as in a common AM broadcasting system, about 20 or less critical-band filters are involved. Speech signals in a frequency range between 50 and 7,000 Hz, for example, can be expressed as power fluctuations in 20 narrow frequency bands. If these power fluctuations are realized by amplitude-modulated band noises in the same frequency bands, a noise-vocoded speech is obtained. This resynthesized speech is almost perfectly intelligible (Ellermeier et al., 2015; Kishida, Nakajima, Ueda & Remijn, 2016) [2,3]. We have been performing a series of analyses and listening experiments from such a viewpoint, but have had no opportunity to give an explanation about how these studies are connected to one another and to previous research. Our present purpose is to indicate how we view our own studies in the process of their development. The chronological order of publications may be broken sometimes in order to describe how we built our thought.

Our first step was to develop a theoretical framework, called Auditory Grammar, to understand how auditory events and auditory streams are formed perceptually (Nakajima, Sasaki, Ueda & Remijn, 2014) [4]. Our basic idea is that an auditory stream, in which auditory events are embedded, is formed perceptually as a linear string of four types of elements (subevents): onsets, offsets (terminations), fillings, and silences. A very simple grammar, in which a filling should be preceded by an onset, and should be succeeded either by another onset or an offset, works here. If an onset cue and an offset cue are detected closely in time and frequency in this temporal order, they are often connected to each other perceptually — even when they do not belong to the same sound physically. This explains two auditory illusions we discovered: the gap transfer illusion and the split-off illusion. If some nonsimultaneous onset cues, of physically different sounds by definition, are close to each other in frequency and time, these onset cues can be connected together perceptually: An illusory melody, i.e., an auditory stream made of illusorily formed auditory events, can be perceived.

2. SPEECH ANALYSES

In Auditory Grammar, if two successive onsets are to be perceived in an auditory stream, there should be at least a filling or a silence in between. Two onsets cannot be directly adjacent to each other in time. An analogy is observed here with a constraint in phonology that two syllabic onsets cannot be directly adjacent to each other. Although the word ‘onset’ is used in a different meaning in the latter case, the idea remains that there cannot be more than one thing to begin one thing. As shown in this example, Auditory Grammar is a phonology of auditory stream formation in general.

In order to develop this idea to understand the mechanisms of speech perception, we analyzed spectral
changes of speech. We thus attempted to find physical correlates of rhythmic aspects of speech, i.e., acoustic patterns showing how syllables appear one by one in speech. Our basic, and yet unpolished, idea was that syllables could be often considered auditory events in speech, and that to find perceptual cues of onsets and offsets in acoustic signals should be important to understand the mechanisms of speech communication. This attempt is still on the way, but we obtained some insight from our analyses of spectral changes of speech. Speech signals in a frequency range up to 7,000 Hz were divided into 20 narrow frequency bands of critical bandwidths. The 20 intensity fluctuations obtained in these frequency bands were subjected to factor analysis. This paradigm was an extension of that created by Plomp’s (1976; 2002) [5,6] research team, in which the spectra of Dutch vowels were analyzed utilizing principal component analysis. The 20 intensity values at each 1-ms time step were considered the variates to be explained by a reduced number of factors. Ueda and Nakajima (2017) [7] analyzed speech signals of eight languages/dialects, and 3- or 4-factor analyses revealed common factors across all these languages.

In the 4-factor analyses, the 4 factors corresponded to 4 frequency ranges, roughly, 50–540, 540–1,700, 1,700–3,300, and above 3,300 Hz. In the 3-factor analyses, one of the factors corresponded to two separate frequency ranges among these ranges: 50–540 and 1,700–3,300 Hz, and the other two factors corresponded to the other two ranges: 540–1,700 and above 3,300 Hz. The above 4 frequency ranges appeared in common in the 4- and the 3-factor analyses, and it seemed that the separation of the whole frequency range into these ranges was vital for speech communication, at least in the eight languages/dialects taken up here, which belong to three different language families. The importance of the 4 frequency ranges for speech communication is in accordance with the well-established fact that 4 carefully chosen frequency bands can make noise-vocoded speech reasonably intelligible (Shannon, 1995) [8]. Furthermore, our own experiment demonstrated that the above 4 frequency ranges could be utilized to generate noise-vocoded speech in German and in Japanese (Ellermeier et al., 2015) [2].

However, the relative cumulative contributions obtained in the above-mentioned 3- or 4-factor analyses were around 40%. This was not sufficient to consider these factors essential for speech communication. We thus decided to resynthesize speech, as noise-vocoded speech, from the obtained factor scores. If a 3-factor analysis preserved essential information for speech communication, then the noise-vocoded speech resynthesized from the factor scores of the 3 factors should have been reasonably intelligible, despite the limited relative cumulative contributions. There was a problem, however. Note that the variates to be analyzed were sound intensities in the critical-band filters. The subspace chosen by a factor analysis generally does not include the acoustically silent point, i.e., the point at which all the intensities are zero. Thus, the resynthesized speech generally lacks silent parts, making it difficult for us to conduct listening tests utilizing such speech stimuli. Kishida et al. (2016) [3], in their study of Japanese speech perception, managed to avoid this problem by modifying the factor analysis so that the acoustically silent point was the origin of the principal component analysis to be performed as the preparatory step for factor analysis. They also smoothed spectra in order to minimize the influence of the harmonic structures observed widely in speech. They thus varied the number of factors from 1 to 9. To increase the number of factors from 2 to 3 increased the intelligibility dramatically from nearly zero to about 70%, and the 4-factor resynthesis made the intelligibility above 80%.

3. ACOUSTIC ANALYSES AND LINGUISTICS

Extracted factors without linguistic interpretation, however, could be just a result of common mathematical procedures. Nakajima, Ueda, Fujimaru, Motomura, and Ohsaka (2017) [9] took up a speech database in which 200 sentences were spoken by three British English speakers. All the spoken phonemes were labeled and related to certain parts of the waveforms in the database, although we had to omit some of them considering our purpose. This enabled us to locate English phonemes into the Cartesian space of the factor scores, obtained in the same way as in Ueda and Nakajima’s (2017) [7] 3-factor analyses. The speech samples of the three speakers were combined and processed together.

These phonemes can be classified into three categories in English phonology: vowels, sonorant consonants, and obstruents (Spencer, 1996) [10]. Vowels, e.g., /æ/, /i/, /u/, or /a/, almost always work as syllable nuclei, and obstruents, e.g., /p/, /b/, /f/, /v/, /s/, or /z/, never can be syllable nuclei. Sonorant consonants, e.g., /w/, /l/, /r/, or /n/, are in between; some of them can be the nuclei of unstressed syllables depending on the contexts. The separation of these three categories was clear in the Cartesian factor space. The likeliness of each phoneme to work as or to come closer to syllable nuclei is called sonority in phonology (Spencer, 1996; Harris, 1994) [10,11], and considered a challenging topic of linguistics (Ball and Müller, 2016) [12]. In the above categorization, vowels are considered to occupy the highest position on the sonority scale, and obstruents the lowest. This was indeed reflected in the Cartesian configuration. The factor score of the factor corresponding to the frequency range 540–1,700 Hz was highly correlated to the sonority. The aperture as described by de Saussure (1959) [13], in
French in 1916, was very close to the sonority. The factor score of the factor corresponding to the frequency range above 3,300 Hz was negatively correlated to this sonority or aperture. This suggests that frequency components above 3,300 Hz can be cues of syllabic boundaries; this has never been stated clearly in linguistics literature to our humble knowledge. A systematic effort to understand sonority and related properties of speech sounds from a psychoacoustic viewpoint is required (Harris, 1994) [11].

The factor corresponding to the frequency range 540–1,700 Hz can be considered an acoustic correlate of sonority as a first approximation, and we will call it the sonority factor. Since sonority is considered vital for syllable formation, it is very likely that it is connected to rhythmic aspects of speech. Yamashita et al. (2013) [14] were interested in the rhythm of infants’ speech, and analyzed the temporal fluctuations of the sonority-factor scores. This factor appeared clearly in infants at 15 months of age, but the frequency range was higher than in adults, corresponding to the smaller sizes of the infants’ articulatory organs. In their analysis of adult Japanese speakers’ speech, the sonority-factor scores often showed periodicities in a range of 0.2–0.4 s. Japanese-learning infants also showed periodicities in this range, but less frequently. When the age increased from 15, to 20 and 24 months, the periodicity in this range became increasingly frequent. This seemed also the case in English-learning infants, but the tendency was less clear and statistically not significant.

4. CONCLUSIONS

Sonority is a subjective or linguistic property of speech sounds closely related to syllable formation. We showed that it was highly correlated to one of the 3 or 4 factors extracted to describe spectral changes of speech, and this factor was closely related to a frequency range around 540–1,700 Hz. The factor scores of this factor were high in vowels, lower in sonorant consonants, and even lower in obstruents in British English. Another factor related to a range above 3,300 Hz was found to be negatively correlated to sonority; the factor scores of this factor were high in obstruents, in which high-frequency components were conspicuous. The factor highly correlated to sonority, the sonority factor, did not appear clearly in 2-factor analyses, but appeared clearly in 3- or 4-factor analyses of all analyzed languages/dialects. The intelligibility of the resynthesized noise-vocoded speech increased broadly monotonically with the number of factors. The intelligibility was nearly zero when the number of factors was 1 or 2, and leaped to around 70% when the number was 3 in a listening experiment in which Japanese sentences were utilized. The intelligibility was above 80% when the number of factors was 4. The sonority factor appeared in speech of 15- to 24-month-old infants, and the factor reflected the speech development of Japanese-learning monolingual infants, in whom periodicities around 0.2–0.4 s as in the Japanese-speaking adults increased with development.

We are particularly interested in how syllables are formed. What is called sonority in phonology plays an important role also in acoustic analyses, and should be investigated from a psychophysical viewpoint.

ACKNOWLEDGMENTS

This study was supported by the Japan Society for the Promotion of Science (Grants-in-Aid for Scientific Research 25242002 and 17H06197).

REFERENCES