AUDITORY STREAM SEGREGATION OF VOWEL SEQUENCE DEPENDING ON SIZE DIFFERENCE: INVESTIGATING THE COHERENCE BOUNDARY AND THE FISSION BOUNDARY

Minoru Tsuzaki, Jóske Nomoto and Chihiro Takeshima
Faculty of Music, Kyoto City University of Arts,
Kutsukake-cho, Oe, Nishikyo-ku, Kyoto-shi, Kyoto 610-1197, Japan
<minoru.tsuzaki@kcua.ac.jp, a-p-borodin@nifty.com, ctakeshima@kcua.ac.jp>

Abstract

A typical stream segregation is demonstrated using the alternation of two frequencies of sinusoids, A and B, in the galloping pattern (ABA-ABA-...). The purpose of the current study was to investigate whether segregation could occur based on the size difference imposed on each vowel in a sequence. The size difference was applied by dilating or shrinking the spectral response of each vocal tract characteristic without changing the fundamental frequencies. Five Japanese vowels, "a, e, i, o, u," were used as the base signals. For each segment of a vowel sequence, the vowel categories were selected randomly from the five types with the restriction that the same category as the two previous should not appear. Two types of boundaries were obtained on the plane of the size difference vs. the alternation speed: one is the coherent boundary; the other is the fission boundary. The results indicate that size-based segregation can occur although it is not as compelling as F0-based segregation.

Auditory stream segregation is defined as the phenomenon by which the human auditory system can perceptually separate out two (or more) sound objects from a mixture of sounds coming from two (or more) different sources. Experimentally, it has been demonstrated by the fact that a sequence of two alternating sounds cannot be heard as a simple alternation, but that it is heard as two parallel, independent repetitions of each sound.

A typical example is the "breaking-down" of a "galloping rhythm" for the repeating ABA-ABA- sequence reported by van Noorden (1975). In one of his experiments, sounds A and B were both sinusoids differing in frequency. When the frequency difference was small and the speed of alternation was slow, the two sounds tended to form a single stream and a rhythmic pattern like a horse's galloping was heard. When the frequency difference was large and the speed of alternation was high, the sounds tended to form two independent streams and a galloping rhythm could not be heard. He also found that there were two boundaries of this perceptual impression depending on the listener's perceptual bias. One was called the coherence boundary, and the other was called the fission boundary. The coherence boundary corresponds to the point where the sequence is segregated into two streams when listeners try to listen to a galloping rhythm. The fission boundary corresponds to the point where a galloping rhythm is inescapably heard when listeners try to single out one of the two sounds.

This auditory characteristic can be regarded as possessing a certain functional value for survival in natural environments. In natural environments, it is very rare to receive a signal coming from a single source. It is mixed with other noises from different sources. It is a reasonable strategy to hear a sound sequence that changes its acoustical property too fast as sounds from two different sources, because a natural vibrating body cannot change its status of vibration so fast.

It is noteworthy that a sinusoid can seldom occur in natural environments. Most sounds are generated by providing a pulse to a resonant body in natural environments. In particular, when animals generate sounds for a certain communication purpose, it is a common strategy to provide pulses periodically like vowels in human vocal communication. When this period becomes shorter than a certain limit, a pitch sensation occurs. As the inverse of the period is the frequency, segregation based on the frequency difference in sinusoids can be regarded as pitch-based segregation. However, the pitch (or the fundamental frequency [F0] in physical terms) is just one of the cues that specify differences in sound sources.

Irino and Patterson (2002) proposed a computational model that could normalize the variation caused by differences in size of resonant bodies. They argued that this size normalization was implemented as a bottom-up process and that the auditory system should be sensitive to detecting the size change of resonant bodies. Following this theoretical study, some experimental evidence has been provided to indicate that the auditory system can extract size information with no learning process, and that it is very sensitive to size difference (Smith et al, 2005; Ives et al., 2005). Tsuzaki et al. (2007) reported a possibility of size-based segregation using an identification task of vowel sequences where the size property was alternated between two values vowel-by-vowel. However, their observation was a rather indirect demonstration. Thus, the purpose of the current paper was to investigate size-based stream segregation more directly using the galloping rhythm paradigm, and to compare it to F0-based stream segregation.

Method

Stimulus and Size Modification

All stimuli were synthesized using a high quality VOCODER system, STRAIGHT (Kawahara et al., 1999), based on samples of five natural Japanese vowels, "a, e, i, o ,u," uttered by male and female speakers. In the frequency domain, an FFT spectrum of a sampled vowel is a multiplication of a continuous spectrum of the vocal tract resonance and a line spectrum of the vocal source periodicity. The STRAIGHT VOCODER can construct a continuous spectrum from the sampled, discrete spectrum by using the F0 adaptive smoothing technique. When the resonant body changes its size proportionally without changing its shape, the frequency response becomes a simple dilated or shrunk copy of the original pattern. When the size reduces, all the formant frequencies become higher while maintaining their mutual ratios; this corresponds to dilation in the frequency region. On the other hand, when the size enlarges, all the formant frequencies become lower; this corresponds to shrinking in the frequency domain as schematically depicted in the top panel of Fig. 1. Thus, one can exclusively change the size property. On the other hand, one can also exclusively change the F0 property (see the bottom panel of Fig. 1).

Although the speakers were instructed to utter in a uniform manner in recording the original vowel tokens, there were slight mismatches in F0s as well as in durations between vowels. To minimize artifacts based on such mismatches, the durations were adjusted manually by trimming and tapering the original tokens on a waveform editor. The F0s were adjusted so that the mean F0 values of each vowel fit the overall mean. Thus, the natural F0 modulation in each vowel was preserved.

To make an ABA-ABA- sequence, the P-center position of each vowel, which could be assumed to correspond to the perceptual time maker in constructing a rhythmic pattern (Marcus, 1981), was calculated based on the temporal power envelope. These P-center points were aligned to the time grid of the rhythmic pattern. Changes in the alternating speed were achieved by shortening each vowel and the pause between them. When shortening the vowels, a couple of frames, which were 1 ms in duration, were dropped to satisfy the intended duration. To avoid shortened vowels from sounding unnatural, the frame dropping was mainly applied to the middle, constant portion of a vowel.

Eleven levels of alternating speed were prepared. Their inter-sound intervals were 75.0, 78.9, 83.3, 88.2, 93.8, 100.0, 107.1, 115.4, 125.0, 136.4, or 150.0 ms. Each sequence lasted about 10 s. Thus, the number of ABA- repetition differed depending on the alternation speed.

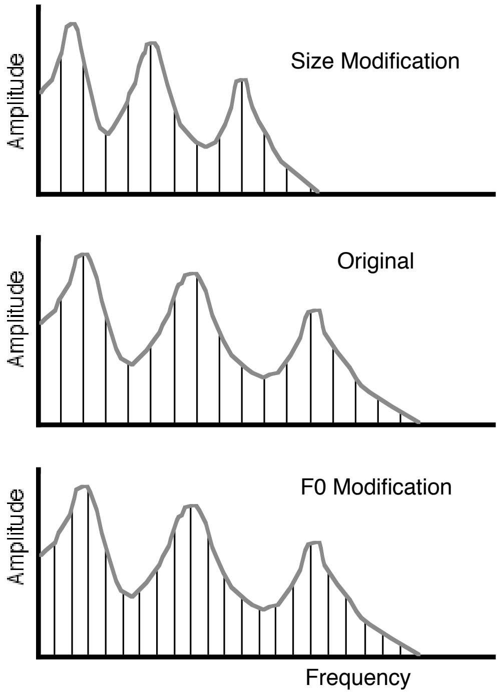

Figure 1. Schematic pictures of two types of vowel modification. The original spectrum is depicted in the middle panel. The top panel depicts the case when the size of a vocal tract is enlarged while maintaining the fundamental frequency. All the formants shift downward proportionally. The bottom panel depicts the case when the fundamental frequency is lowered while maintaining the vocal tract size. The spectral envelope is preserved, while the fundamental frequency is lowered and the spacing between the harmonics becomes tight.

Both the size and F0 changes can be described in terms of ratios of modification to the base pattern. Ten levels of the separation between the alternating sounds were prepared. They ranged from 0 to 18 semitones by a 2-semitone step. Vowel categories were randomly selected from five, i.e., "a, e, i, o, u," with the restriction that the same vowel category never repeated within three slots in succession.

Procedure and Participants

Participants were required to estimate how prominent was the impression of the galloping rhythm on a five-point scale (5 for the most prominent; 1 for the least prominent) for each combination of alternation speed (11 levels) and sound separation (10 levels) in 16 listening conditions. The 16 listening conditions consisted of a factorial combination of four factors as follows: (1) listening bias; (2) operated stimulus parameter; (3) original voice source; (4) order of ABA pattern.

The listening bias factor corresponded to whether participants were instructed to try to hear a galloping rhythm as possible, or to try to single out one of the two sounds. The operated stimulus parameter factor corresponded to whether the sound separation between A and B was realized by the size difference or by the F0 difference. The original voice source factor corresponded to whether the stimuli were synthesized based on the male voice or the female voice. The ABA order factor corresponded to which of two sounds, a high sound or a low sound, was assigned to A or B, respectively. In the case of size difference, high sound means sounds with high formant frequencies, i.e., sounds coming from a small resonant body. In the case of F0 difference, high sound means sounds with a higher pitch.

All stimuli were synthesized off-line. The experimental sequence including the response recording was managed with a computer (Apple iMac G5) that also controlled a DSP interface (Capybara 320 + Kyma 5.0, Symbolic Sound co.) for D-to-A conversion with 16-bit quantization, 44-kHz sampling rate. The stimuli were presented diotically through headphones (Sennheiser HD600) amplified with a headphone amplifier (Luxman P-1). Participants were tested in a sound-treated room.

Three participants participated in the experiment. One of them was the second author of the current paper, and he repeated all the combinations twice. The other two gave a judgment on each of the combinations once.

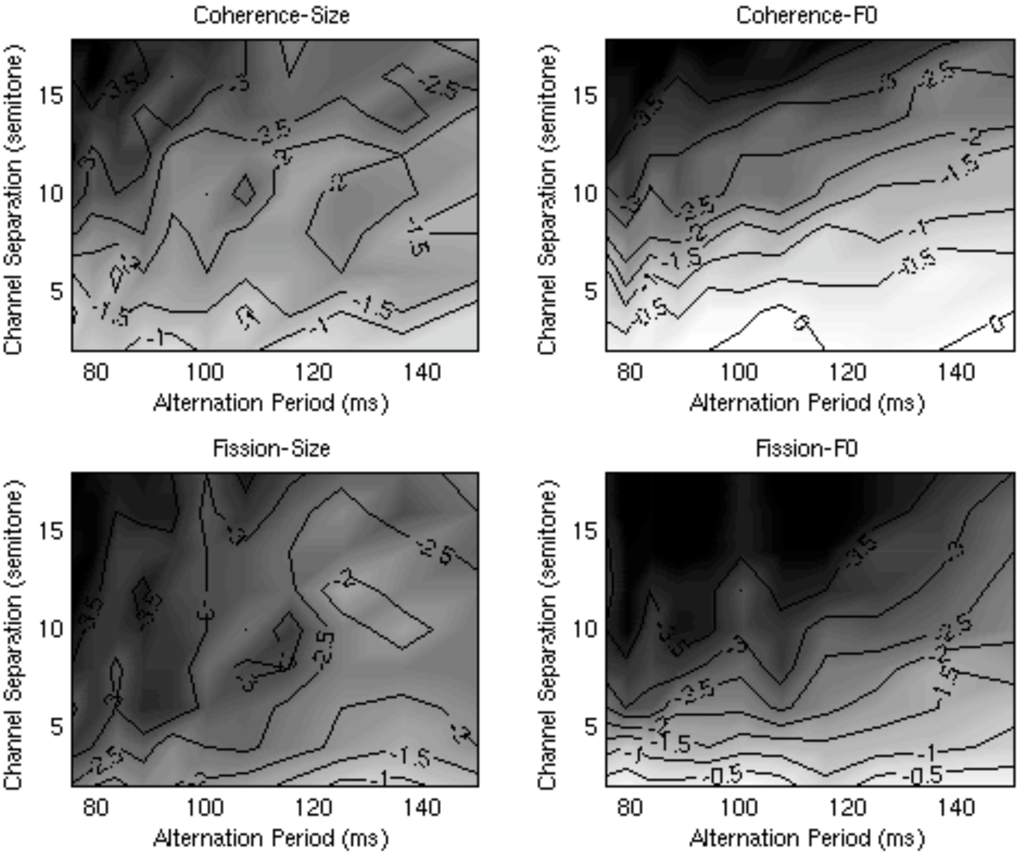

Figure 2. Contour plots of the "equal-streaming" levels. Panel (a) depicts the results when listeners were instructed to hear sequences coherent under size modulation; panel (b) depicts the results when listeners were instructed to segregate sequences under size modulation; panel (c) depicts the results when listeners were instructed to hear sequences coherent under F0 modulation; panel (d) depicts the results when the listeners were instructed to segregate sequences under F0 modulation.

Results and Discussion

Medians of rating score were obtained for each of 110 points combining 11 alternation speeds and 10 parameter differences for each combination of the listening bias and the operated stimulus parameter by pooling over the original voice source, the ABA order, and the participant. For each alternation speed, the sequence whose parameter difference was at the reference level, i.e., 0, can be assumed to be a control condition. The median scores of these control conditions were subtracted from the other corresponding scores. The more negative the score, the more prominent the segregation impression. In Fig. 2, Equal-segregation contours are plotted by interpolating the values for these 99 points. The gray scale represents the degree of segregation prominence. The upper panels depict the results of the condition where the participants tried to hear a galloping rhythm, corresponding to the coherence boundary. The lower panels depict the results of the condition where the participants tried to single out one of the two sounds, corresponding to the fission boundary. The panels in the left column depict the results based on the size difference. The panels in the right column depict the results based on the F0 difference.

In the region above the coherence boundary, a sequence is irresistibly segregated into two streams whatever the listeners' perceptual bias is. Bregman (1990) called this type of segregation the primary auditory segregation. If the line corresponding to –3.0 is assumed to represent a coherence boundary, the F0-based segregation case (the top, right panel of Fig. 2) showed the interaction between the separation degree and the alternation speed as shown in the study by van Noorden (1975). The primary segregation was likely to occur with a smaller F0 separation as the alternation speed was faster. The interaction, however, seemed to be much weaker than van Noorden's data. Compared to the results of F0-based segregation, the region of the primary segregation by size difference appeared to be limited (the top, left panel of Fig. 2), while a similar interaction between the separation degree and the alternation speed was observed.

If the fission boundary, below which the galloping rhythm is irresistibly heard, is defined as the line corresponding to –0.5, it is observable in the F0-based condition (the bottom, right panel), while it is missing in the size-based condition (the bottom, left panel). While F0 is a more robust cue than size, it can intentionally be controllable with a certain flexibility by a speaker. In contrast, the size of the vocal tract cannot be changed easily by intention. Therefore, the current results suggest that the auditory system can use the size difference to single out one sound from the other with less difference than it does with the F0 difference.

One can argue that the current finding is simply a demonstration of stream segregation based on timbre differences (Hartmann & Johnson, 1991), and that the results can be explained parsimoniously by the concept of timbre without mentioning the size concept. This argument could be valid if the concept of timbre were clearly defined. However, this is not the case. When one refers to timbre difference, it is simply saying that the sounds are different from each other although their pitch and loudness are identical. Generally, it is assumed that timbre is represented in a multi-dimensional space. This means that the concept of timbre is still ambiguous. It is not anomalous to assume that the size-related aspect composes one such dimension. The merit of the size concept is that it is a grading system like pitch and loudness.

Acknowledgement

This work was supported by the Grant-in-Aid for Scientific Research (C), No. 17530529, Japan Society for the Promotion of Science.

References

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Massachusetts: MIT Press.

Irino, T. & Patterson, R. (2002). Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform. *Speech Communication*, 36, 181-203.

Ives, D. T., Smith, D. R. R. & Patterson, R. D. (2005). Discrimination of speaker size from syllable phrases. *Journal of Acoustical Society of America*, 118, 3816-3822.

Hartmann, W. M. & Johnson, D. (1991). Stream segregation and peripheral channeling. *Music Perception*, 9(2), 155-184.

Kawahara, H., Masuda-Katsuse, I. & de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27, 187-207.

Marcus, S. M. (1981). Acoustic determinants of perceptual center (P-center) location. *Perception & Psychophysics*, 30, 247-256.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H. & Irino, T. (2005). The processing and perception of size information in speech sounds. *Journal of Acoustical Society of America*, 117, 305-318.

Tsuzaki, M., Takeshima, C., Irino, T. & Patterson, R. D. (2007). Auditory stream segregation based on speaker size, and identification of size-modulated vowels sequences. In B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp & J. Verhey (Eds.), *Hearing - from basic research to applications*. Heidelberg: Springer Verlag.

van Noorden, L. P. A. S. (1975). *Temporal Coherence in the Perception of Tone Sequences*. Unpublished doctoral dissertation, Eindhoven University of Technology.