Perceptual organization of onsets and offsets of sounds

Nakajima, Yoshitaka
Department of Acoustic Design, Kyushu University

Sasaki, Takayuki
Miyagi Gakuin Women's College

Remijn, Gerard B.
Department of Visual Communication Design, Kyushu University

Ueda, Kazuo
Department of Applied Information and Communication Sciences, Kyushu University

出版情報：Journal of Physiological Anthropology and Applied Human Science. 23 (6), pp.345-349, 2004-12-10. 日本生理人類学会
バージョン：
権利関係：

Journal of
PHYSIOLOGICAL
ANTHROPOLOGY
and Applied Human Science

Perceptual Organization of Onsets and Offsets of Sounds

Yoshitaka Nakajima[1)], Takayuki Sasaki[2)], Gerard B. Remijn[3)] and Kazuo Ueda[4)]

1) Department of Acoustic Design, Kyushu University
2) Miyagi Gakuin Women's College
3) Department of Visual Communication Design, Kyushu University
4) Department of Applied Information and Communication Sciences, Kyushu University

Abstract Several illusory phenomena in auditory perception are accounted for by using the event construction model presented by Nakajima et al. (2000) in order to explain the gap transfer illusion. This model assumes that onsets and offsets of sounds are detected perceptually as if they were independent auditory elements. They are connected to one another according to the proximity principle to constitute auditory events. This model seems to contribute to a general cross-modal theory of perception where the idea of edge integration plays an important role. Potential directions in which we can connect the present paradigm with speech perception are indicated, and possibilities to improve artificial auditory environments are suggested. *J Physiol Anthropol Appl Human Sci 23(6): 345–349, 2004* http://www.jstage.jst.go.jp/browse/jpa

Keywords: auditory organization, onsets, offsets, auditory events, the gap transfer illusion, the proximity principle, the event construction model

Introduction

It is important to investigate how we listen to sound patterns in order to improve the quality of our life, because auditory communication is vital for maintaining human society. There are many problems to be solved in current acoustic environments. In an airport, we often have to catch the flight number of our plane in a very noisy situation, sometimes in a foreign language. Mobile phones, whose sound quality is not necessarily very good, are very important tools for organizing our social behavior. Radio programs are convenient means for many people to enrich their intellectual life, and seem to be increasingly important as the Internet, which still has a lot of technical limitations, expands. The number of senior citizens is unprecedentedly large in many countries, and it is inevitable that they are also involved in the busy auditory communication of the modern society, often with impaired hearing.

In order to improve acoustic environments in the modernized world, where many different people communicate with each other in many different ways, it is not sufficient to make acoustic signals simply easier to detect. Sounds used for communication should not be too loud, and, simultaneously, should be clear enough even in noise. The first thing to try is to reduce environmental noise, but this is often difficult especially in an urban environment. It is necessary to find a way to make acoustic signals, especially speech signals from loudspeakers or telephones, clearer without making them too loud. For this goal, we have to accumulate data on how we perceive the temporal structures of sounds.

Characteristics of Auditory Organization

Principles of auditory organization have been investigated systematically for a few decades (Handel, 1989; Bregman, 1990; Warren, 1999). Researchers agree with each other that the concept of auditory streams is vital in our attempts to understand auditory organization. Acoustic information basically reflects movements and activities of objects and organisms that are close to the perceiver or of the perceiver itself. In vision, information from an object drawing the viewer's attention usually causes excitation of a compact area on each retina at each moment, and this makes the job of the brain to extract the information of this object easier. One of the reasons that make this possible is that objects to be attended to are often situated near the center of the visual field, and that the rays from the objects to the retinae are not mixed with rays from other objects. The situation is quite different in audition. Information of one object, or one sound source, is often scattered on the basilar membranes, and in most cases sounds from several sources are mixed together. One may think that the brain suppresses unnecessary signals. This may be the case, but it does not mean that some signals are discarded. Because one of the most important functions of the auditory system is to detect dangers and changes in the environment, to discard seemingly unimportant signals does not necessarily mean that the most efficient use of the information given to the ear is made. We may notice that the sound of rain suddenly

weakened even though we have not attended to it at all. An interesting point is that we are often able to "remember" how the rain had sounded before the change. As this example shows, auditory signals seem to be processed somehow whether attended to or not.

Sometimes, auditory signals to which we pay attention are considered to be organized as figures, and the other signals as a ground. This kind of idea comes from an analogy between audition and vision. Although the analogy may be convenient as a first step to understand auditory organization, it may be misleading because auditory signals that are not attended to are often organized with a clear structure. A ground in visual organization cannot have a clear structure by definition, but, when a melody and an accompaniment are perceived, both of them have clear rhythmic structures that can be in conflict with each other in some cases. It is inappropriate to argue that only one of them has a clear shape (Bregman, 1990). A promising way to solve the problem is to introduce the concept of auditory streams instead of the concepts of figure and ground.

The majority of the researchers investigating auditory perception seem to accept the idea that auditory streams play important roles in auditory organization both in experimental rooms and in our everyday life. Auditory streams are indeed similar to figures in vision, but a great difference is that auditory streams may stay behind other auditory streams while figures always remain near the center or the front of the subjective world. We sometimes begin to pay attention to an auditory stream (for example, chirps of a cricket or the noise of a hard disk drive) noticing that it has been ongoing in the auditory background for some time.

An auditory stream is a linear string of auditory events and silences. A phrase in speech, a melody in music, a sequence of footsteps, and a continuous fan noise are all typical examples of auditory streams. Strictly speaking, the percepts of these sequences are auditory streams. Our ears are often exposed to a disorganized flood of sound energy, and to distinguish auditory streams is an important step in constructing the subjective world of sound. The definition of an auditory event can diverge among researchers and contexts. We will use this word in a strictly basic sense in this report. A syllable in speech, a note in music, and a single footstep are typical auditory events. In some cases, a consonant at the end of a syllable, a short spoken word consisting of a few syllables, an attack of a musical note, or a note together with grace notes can be a single auditory event. The definition may fluctuate a little.

The gap transfer illusion

In the following, we would like to clarify how the concept of auditory events works through empirical research. Nakajima and Sasaki (1993) found a new auditory illusion, which was reported in English later (Nakajima et al., 2000, Fig. 1). We generated stimulus patterns where a long ascending frequency glide of 1500 ms, from 640 to 2160 Hz, and a short descending frequency glide of 500 ms, from 1440 to 960 Hz, crossed each other, at 1175.8 Hz, at their temporal centers. These glides moved at the same speed in logarithmic frequency in opposite directions. They were at the same level, and their rise and fall times were 10 ms. When the short glide had a temporal gap of 100 ms in the middle, and the long glide was continuous, the perception of this pattern was veridical. A typical percept was a long ascending glide accompanied by a short descending glide with a temporal gap in the middle. This may not be a very impressive result, but it should be noted that bouncing perceptual components, which often appear in this kind of stimulus patterns of crossing frequency glides (Halpern, 1977; Bregman, 1990), rarely appeared in the present case.

When a temporal gap of the same duration was introduced into the middle of the long glide instead of the short glide, the typical percept remained basically the same. Although the physical temporal gap was in the long ascending glide, it was perceived, surprisingly, as if it had still been in the short glide. This phenomenon is what we call 'the gap transfer illusion.' The same kind of illusory transfer of a temporal gap took place also when this stimulus pattern was reversed in time.

Tsunashima and Nakajima (2002) used harmonic frequency glides consisting of three components, and showed that the illusion took place robustly also in their paradigm, unless there were amplitude differences between corresponding components of the crossing glides. They also attempted to relate the temporal gap perceived in the gap transfer illusion to the perception of stop consonants in speech. They demonstrated a stimulus pattern in which a long harmonic frequency glide of 4000 ms, from 126 to 317 Hz, and a short harmonic frequency glide of 400 ms, from 209 to 191 Hz, crossed each other, at 200 Hz, at their temporal centers. These glides had three formants each to be perceived as vowels /a/, and moved at the same speed in logarithmic frequency in opposite directions. The long glide had a temporal gap of 100 ms in the middle, and a very short noise, constituting the beginning of a consonant /k/, was placed in the middle of this gap. A formant transition for /k/ was also introduced immediately after the gap. If the long glide with the temporal gap and the very short noise had been presented alone, a stop consonant /k/ preceded and succeeded by rather long vowels /a/ would have been perceived. Note that the temporal gap must have been crucial to cause the perception of /k/. When the short frequency glide was added, however, the consonant /k/ or sometimes a different stop consonant was perceived not in the long component, but in the short component. The short component included no physical clue of a stop consonant except for the very short noise, whose physical allocation was undetermined. The short component, which would have been perceived as a single vowel /a/ when presented alone, was perceived typically as /aka/ in the present condition. The most plausible explanation is that the temporal gap in the long component was transferred into the short component subjectively, and this gap was utilized by the perceptual system as a clue for /k/. Strictly speaking, the gaps before and after the

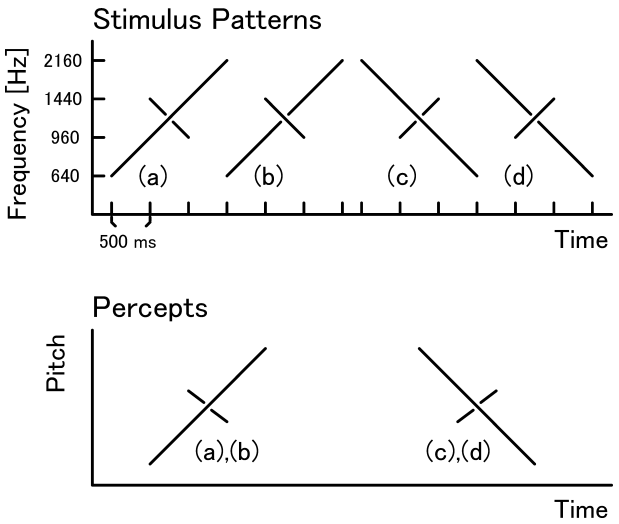

Fig. 1 The gap transfer illusion.

noise must have been utilized as clues. An interesting point is that the temporal gap itself often did not appear in the percept of this pattern when the transferred stop consonant was perceived.

For a theoretical examination, we take up the very first example where frequency glides of 500 and 1500 ms cross each other. Suppose that the ascending glide, either the shorter or the longer of the two components, has a temporal gap in the middle. If this glide is shorter (500 ms), the gap is perceived veridically as being in this glide [Fig. 1(c)]. If this glide is longer (1500 ms), however, the gap is perceived as being in the other glide [Fig. 1(b)]. Because these two stimulus patterns are almost the same within the 500-ms range around the crossing point, what happens more than 200 ms before or after the temporal gap must affect whether the gap is perceived in the veridical or in the illusory position.

The event construction model

We proposed a simple model, 'the event construction model,' in order to understand the mechanism of the gap transfer illusion (Nakajima et al., 2000; see also Nakajima and Sasaki, 1996). The basic idea is that onsets and offsets detected by the auditory system behave as if they were independent 'subevents.' A temporal gap is accompanied by, or includes, an offset of a sound and an onset of another sound in this order. If the offset of a sound at the beginning of the gap is connected perceptually to an onset which precedes it, then we get a new 'auditory event.' If the onset at the end of the gap and an offset are connected perceptually in this order, then we get another auditory event.

Probably, an onset and an offset can be connected with each other more easily when they are closer to each other in time and frequency, because then they are more likely to interact in the perceptual mechanism. This means that the onset and the offset of the shorter glide can be connected more easily to the offset and the onset bounding the temporal gap, compared with their counterparts of the longer glide. In the gap transfer illusion, the model explains that the gap is more likely to be perceived in the shorter glide. We have introduced one of the Gestalt principles, the proximity principle, in order to understand the mechanism of the illusion.

Nakajima et al. (2000) reported a closely related phenomenon. They used a stimulus pattern where a long ascending glide of 2500 ms, from 421.7 to 2371.3 Hz, and a short descending glide of 500 ms, from 1188.5 to 841.4 Hz, crossed each other at 1000 Hz in opposite phases. Thus, the temporal envelope of the whole stimulus pattern had a gap of 38 ms in the middle (if the temporal boundaries of a gap are assumed to be located at the points at which the sound pressure level is 3 dB below the maximum level). There was no particular reason to determine to which glide the gap should belong physically, but the general tendency among the observers was that clearer discontinuity was perceived in the shorter glide. This perspective was in line with the idea that the auditory system is more likely to allocate temporal gaps to shorter, than to longer, tones.

In order to confirm the validity of the event construction model, Remijn and Nakajima (2000) made stimulus patterns of two crossing glides. In one of the patterns, a short glide of 500 ms descended from 1783.8 to 1261.3 Hz, and a long glide of 2000 ms ascended from 750 to 3000 Hz. These glides crossed each other at 1500 Hz at their central positions, and had a common temporal gap (a completely silent period) of 20 ms at the crossing point. The rise and fall times throughout the pattern were 20 ms. The discontinuity caused by the common temporal gap was perceived mainly in the short glide, and the long glide was perceived as reasonably continuous. The long glide with the gap was perceived as more continuous than when presented alone. That is, adding a short discontinuous glide made the longer, discontinuous glide more continuous perceptually.

The event construction model explains the result in the following way. The offsets of glides at the beginning of the common gap are detected as a single offset because they are simultaneous and close to each other in frequency. This offset is perceptually connected with the onset at the beginning of the short glide, which is closer in time and frequency than the onset at the beginning of the long glide. This perceptual connection of an onset and an offset makes a percept of a short tone corresponding roughly to the first half of the short glide. The onsets of glides at the end of the gap are detected as a single onset, and this onset is perceptually connected with the offset at the end of the short glide, thus making a percept of a short tone corresponding roughly to the last half of the short glide. Because the offset and the onset bounding the gap have been interpreted, the perceptual system need not interpret them again. If the gap is not long enough to give a perceptual clue for a silent period corresponding to an absence of sound energy, there is no reason that the long glide be perceived as discontinuous.

This model enabled Nakajima et al. (2000), Remijn et al. (2001), and Remijn and Nakajima (2004) to predict and establish a new auditory phenomenon, and, thus, proved its own feasibility. Two partly overlapping frequency glides could be perceived as consisting of a long glide, corresponding roughly to the duration of the whole pattern, and a short tone, corresponding roughly to the duration of the overlap. For example, a glide moved for 1400 ms from 367.2 to 965.7 Hz, and another glide of the same duration began 1200 ms later and moved from 1035.5 to 2723.7 Hz. They overlapped each other for 200 ms, keeping a distance of 0.3 octave. The rise time and the fall time bounding the overlap were 4 ms. A typical percept of this pattern was a long ascending glide corresponding roughly to the duration of the whole pattern accompanied by a short tone corresponding roughly to the duration of the overlap. The event construction model predicted the perception of the short tone in a simple manner: the onset of the second glide and the offset of the first glide were close to each other in time and frequency, and they would be connected perceptually to construct an auditory event, i.e., the short tone. To explain the perception of the long glide is not very easy, and requires more data on the perceptual continuity of glide tones (Ciocca and Bregman, 1987; Remijn et al., 2001).

New directions

It will be productive in the future to give a wider view to our idea, which features the perceptual integration of auditory onsets and offsets that are close to each other in time and frequency. It has been known for a long time that neurons in the auditory cortex can react to acoustic onsets, offsets, or both (e.g., Whitfield and Evans, 1965), and recent efforts concentrate on finding neurophysiological evidence for the integration of temporal edges of auditory signals (Philips et al., 2002). DeCharms et al. (1998) suggested that processes behind the integration of auditory stimulus edges at neural levels resemble processes behind the integration of visual stimulus edges. The event construction model can be firmly related to the idea of feature integration in vision, in which the Gestalt principle of proximity also plays an important role. The idea of feature integration suggests that a visual scene is decomposed into object features, such as edges or conjunctions of edges, and that similarity and proximity between these features facilitate their perceptual integration. A number of recent studies have also provided psychophysical evidence for the idea of edge integration in vision (Barlow, 1999; Geisler et al., 2001). Our findings with regard to the perceptual integration of onsets and offsets of sounds, therefore, are in a position to contribute to a general cross-modal theory of perception.

Now we are trying to find phenomena that could connect our in vitro paradigm with realistic situations (Ueda et al., 2004). Wang and Nakajima (2004) succeeded in generating auditory stimulus patterns where illusory recouplings of onsets and offsets seem to cause the perception of Chinese syllables. If this kind of illusory percept with a clear temporal structure could connect underlying mechanisms of auditory event perception and phonological rules or constraints of different languages, we would be able to find efficient ways of auditory communication in modern society, which is heavily dependent on artificial environments. For example, we could develop broadcasting systems, telephone systems, or hearing aids that are friendlier to senior citizens and hearing-impaired people, and we could also improve audio-guidance systems for blind people (Iwamiya et al., 2004). One possible application of our model is a system which enhances onsets and offsets of syllables in speech in order to increase intelligibility without increasing loudness.

Conclusions

The event construction model, in which the perceptual connection of onsets and offsets plays an important role, revealed its explanatory power for some illusory auditory phenomena and its utility to find new auditory phenomena that could be connected to our linguistic behavior. We are in a new field of research, where underlying mechanisms of auditory perception and phonological rules or constraints in languages should be understood from a unified viewpoint. One of our immediate goals is to find more efficient ways of auditory communication, especially in artificial auditory environments.

Acknowledgments We are grateful to Gert ten Hoopen, Koji Otsuka, and Jonathan Goodacre for their valuable comments and suggestions. A part of this paper was presented by the first author in ‘The Forum Acusticum Sevilla 2002.’ The present study was supported by the 21st Century Center of Excellence (COE) Program of the Japanese government entitled ‘Design of Artificial Environments on the Basis of Human Sensibility.’ and a Grant-in-Aid for Scientific Research provided by JSPS (14101001 in the fiscal years 2002–2004).

References

Barlow HB (1999) Feature detectors. In: Wilson RA, Keil FC eds. The MIT Encyclopedia of the Cognitive Sciences. MIT Press, Cambridge

Bregman AS (1990) Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, Cambridge

Ciocca V, Bregman AS (1987) Perceived continuity of gliding and steady-state tones through interrupting noise. Percept Psychophys 42: 476–484

DeCharms RC, Blake DT, Merzenich MM (1998) Optimizing sound features for cortical neurons. Science 280: 1439–1443

Geisler WS, Perry JS, Super BJ, Gallogly DP (2001) Edge co-occurrence in natural images predicts contour grouping performance. Vision Res 41: 711–724

Halpern L (1977) The effect of harmonic ratio relationships on auditory stream segregation. McGill University, Psychology Department (unpublished research report)

Handel S (1989) Listening: An Introduction to the Perception

of Auditory Events. MIT Press, Cambridge

Iwamiya S, Yamauchi K, Shiraishi K, Takada M, Sato M (2004) Design specifications of audio-guidance systems for the blind in public (in this issue)

Nakajima Y, Sasaki T (1993) Perceptual transfer of onsets and offsets between crossing glide tone components. IEICE Technical Report SP92-146: 73–80 [*In Japanese*]

Nakajima Y, Sasaki T (1996) A simple grammar of auditory stream formation (abstract). J Acoust Soc Am 100: 2681

Nakajima Y, Sasaki T, Kanafuka K, Miyamoto A, Remijn G, ten Hoopen G (2000) Illusory recouplings of onsets and terminations of glide tone components. Percept Psychophys 62: 1413–1425

Phillips DP, Hall SE, Boehnke SE (2002) Central auditory onset responses, and temporal asymmetries in auditory perception. Hear Res 167: 192–205

Remijn GB, Nakajima Y (2000) The perception of two crossing frequency glides sharing a very short temporal gap. Rep Acoust Soc Jpn H-2000-107: 1–8

Remijn GB, Nakajima Y (2004) The perceptual integration of auditory stimulus edges: An illusory short tone in stimulus patterns consisting of two partly overlapping glides. J Exp Psychol Hum Percept and Perform (in press)

Remijn GB, Nakajima Y, ten Hoopen G (2001) Continuity perception in stimulus patterns consisting of two partly overlapping frequency glides. J Music Perception and Cognition 7: 77–91

Tsunashima S, Nakajima Y (2002) Demonstrations of the gap transfer illusion. Proc of the 7th International Conference on Music Perception and Cognition, Sydney

Ueda K, Nakajima Y, Akahane-Yamada R (2004) An artificial environment is often a noisy environment: Auditory scene analysis and speech perception in noise. J Physiol Anthropol Appl Human Sci (in press)

Wang H, Nakajima Y (2004) Illusory Chinese syllables perceived in stimulus patterns consisting of two partly overlapping harmonic glides. Trans Tech Comm Psychol Physiol Acoust, Acoust Soc Jpn 34: 553–556

Warren RM (1999) Auditory Perception: A New Analysis and Synthesis. Cambridge University Press, Cambridge

Whitfield IC, Evans EF (1965) Responses of auditory cortical neurons to stimuli of changing frequency. J Neurophysiol 28: 655–672

Received: September 6, 2004

Accepted: October 5, 2004

Correspondence to: Yoshitaka Nakajima, Department of Acoustic Design, Kyushu University, 4–9–1 Shiobaru, Minami-ku, Fukuoka 815–8540, Japan

Phone: +81–92–553–4558

Fax: +81–92–553–4520

e-mail: nakajima@design.kyushu-u.ac.jp