An Artificial Environment is Often a Noisy Environment: Auditory Scene Analysis and Speech Perception in Noise

Kazuo Ueda1, Yoshitaka Nakajima2 and Reiko Akahane-Yamada3

1) Department of Applied Information and Communication Sciences, Kyushu University
2) Department of Acoustic Design, Kyushu University
3) ATR Human Information Science Laboratories

Abstract Our auditory system has to organize and to pick up a target sound with many components, sometimes rejecting irrelevant sound components, but sometimes forming multiple streams including the target stream. This situation is well described with the concept of auditory scene analysis. Research on speech perception in noise is closely related to auditory scene analysis. This paper briefly reviews the concept of auditory scene analysis and previous and ongoing research on speech perception in noise, and discusses the future direction of research. Further experimental investigations are needed to understand our perceptual mechanisms better. J Physiol Anthropol Appl Human Sci 24(1): 129–133. 2005 http://www.jstage.jst.go.jp/browse/jpa

Keywords: second language acquisition, signal-to-noise ratio, cross-modal covariances

Introduction

An artificial environment may include or accompany many noise sources, such as airplanes, cars, ambulances, air conditioning systems, doors, televisions, computers, a whistle-kettle on a gas range, phones ringing, and other people talking. However, our auditory system somehow succeeds, in most of the cases, to pick up a target sound out of such a noisy environment, and this capability is supposed to stem from our basic perceptual mechanisms. The situation is named as auditory scene analysis (Bregman, 1990). Auditory scene analysis has become one of the key concepts of modern auditory psychology. Meanwhile, research on speech perception in noise has 50 years of history, and is closely related to auditory scene analysis.

This paper consists of four main parts: a brief introduction to auditory scene analysis, an overview of previous research on speech perception in noise, an overview of previous and ongoing research on English consonant perception and training, and discussion about the future direction of research.

What is Auditory Scene Analysis?

Bregman (1990) proposed the concept of auditory scene analysis that provided a research framework from a Gestalt psychological perspective and a kind of ecological perspective. The concept is partly characterized by its basic assumptions: The task for the human auditory system is to organize, or to segregate, a target sound from the mixture of many components of sounds in our daily life, without any help from other sensory modalities except from the information already stored in long-term memory. The concept also imposes a constraint that the owner of the auditory system is not allowed to move around. Even in such a severely constrained situation, we can get a nice auditory picture of an outer world. It is an amazing capability, and Bregman focuses on how our auditory system accomplishes the task.

He proposed to use the concept of “auditory streams.” An auditory stream means “our perceptual grouping of the parts of the neural spectrogram that go together” (Bregman, 1990, p. 9). Moreover, he pointed out that our auditory system realizes two kinds of integration: sequential integration and simultaneous integration. The sequential integration means “the process that connects events that have arisen at different times from the same source” (Bregman, 1990, p. 30). The simultaneous integration means to integrate spectral components that occur simultaneously in separate regions of a spectrum. Note that, however, these two kinds of integration do not thoroughly cover what our auditory system can integrate: There is some evidence showing that an integration of nonsimultaneous spectral components in separate regions of a spectrum can occur (Nakajima et al., 2004).

Gestalt principles of perceptual grouping, such as the principle of proximity, the principle of similarity, the principle of continuity, and so on, can be regarded as rules for auditory scene analysis. These Gestalt principles form the basis of auditory scene analysis. However, Bregman thinks that the principles are too abstract, and that the principles do not take into account differences between light and sound. For example,
Speech Perception in Noise

sounds can bend around obstacles easily and the information conveyed through sounds is well preserved even after the bending, but this is not the case for light. These characteristics are reflected in the ways our sensory systems utilize those stimuli: the systems often utilize lights to know the shapes of objects, but sound to detect clues related to the existence of objects. This way of thinking, which originates from ecology, is one of the features of auditory scene analysis research.

Our auditory system needs some built-in constraints to solve the problem, i.e., to make proper integration or segregation of sounds, based on incomplete information. Bregman pointed out that the built-in constraints should be divided into two categories: unlearned constraints and learned constraints. Unlearned constraints, which realize primitive segregation, are built in from our birth, and need no conscious effort of attention, whereas learned constraints, which realize schema-based segregation, are built after our birth, and need conscious effort of attention. Our skill to perceive and to segregate speech is a typical example of schema-based segregation. Bregman proposed a two-part theory of perceptual organization that consists of a first process, i.e., primitive auditory scene analysis, and a second process, i.e., a description-building process using schema. This theory gives an alternative explanation of duplex perception of speech, in place of the motor theory of speech perception proposed by Liberman (1982).

Previous Research on Speech Perception in Noise

Bregman's work was quite valuable in directing the attention of researchers to the perceptual organization of sound. Concerning speech perception, Bregman suggested that even an infant had to analyze the auditory scene: to distinguish a cradle's squeak and her mother's speech, auditory scene analysis was needed. Bregman supposed that this capacity should come from innately given constraints; otherwise, the infant could not start learning speech.

Prior to Bregman's work, there had been a tradition of research concerning speech perception in a noisy environment. For example, Miller and Nicely (1955) investigated the intelligibility of English consonants in noise, and revealed that the extent of noise tolerance was different from consonant to consonant. The accuracy of word recognition depended on predictability provided by the sentence that embeds the word (Miller et al., 1951; Miller, 1962; Duffy and Golas, 1974). The familiarity of words was another variable that affected word recognition in noise (Pollack et al., 1959). The amount of masking effect depended on the kinds of masker noise: A babble noise was more effective in masking speech sounds than a steady random noise, and increasing the number of the talkers in the babble noise made the masking more effective (Carhart et al., 1975).

Non-native listeners find it harder to hear speech sounds in noise than native listeners do. A study by Buus and his colleagues (Buus et al., 1986) determined the sound pressure levels of the background white noise at which participants could repeat correctly about 50% of simple American-English sentences. French-native listeners were less tolerant of background noise than American-English-native listeners. This tendency was even more pronounced if the French-native listeners had fewer opportunities to talk in English. A similar investigation was conducted with fluent non-native speakers of English having various language backgrounds (Florentine, 1985; Mayo et al., 1997). The Speech Perception in Noise (SPIN) test (Kalikow et al., 1977) and a babble noise were used to measure noise-tolerance levels. A target word of the SPIN test was always a monosyllabic noun presented at the end of a test sentence. Half of the test sentences had highly predictable targets, whereas the other half did not. The noise tolerance levels as well as the slopes of the psychometric functions of the noise tolerance levels were determined to clarify the effect of noise. The non-native listeners seemed to have a disadvantage in utilizing contextual cues in sentences. The earlier the age of starting second-language acquisition was, the less the disadvantage they suffered. However, bilinguals never performed like monolinguals in noise. The disadvantage of non-native listeners was also shown in the recognition of Japanese short sentences (Hatoh et al., 1998).

English Consonant Perception and Training

Although the recent research on word recognition in noise seems to emphasize the importance of predictability given by a carrying sentence, it does not mean that phoneme recognition is unimportant. Contextual cues conveyed through a carrying sentence would originate from some easily detectable words, and the evoked cues would involve previous knowledge that relates to a target word. However, if one cannot detect any words in a sentence, it is obvious that one cannot form a context from the sentence. Then, one would never be able to guess a target word from the context. Thus, it is still important to shed light on the problem of speech perception in noise under the condition that words are presented in isolation.

Disadvantages for non-native listeners have been observed also in consonant perception. Adding reverberation or babble noise more greatly impairs non-native listeners' perception of English consonants than that of native listeners (Nábělek and Donahue, 1984; Takata and Nábělek, 1990).

American-English consonant perception in white noise by native- and non-native listeners has been investigated in the first and the third authors' laboratories (e.g., Ueda et al., 2002). The focus of our study is on the identification accuracy of American-English /r/ vs. /l/ in an isolated word with systematically varied signal-to-noise ratios. A minimal pair of English words contrasting in /r/ and /l/ was displayed in English orthography on a computer screen. Simultaneously, one of the words in the pair was played over headphones. The participants chose the target word on the screen in a two-alternative forced choice paradigm. The experiment is still going on. Current results are shown in Fig. 1. These results
show that (a) identification accuracy by American-English listeners was nearly 100% in the no-noise condition, whereas it was 70% at a signal-to-noise ratio of -21 dB, and (b) the Japanese listeners and the American-English listeners showed similar identification accuracy for the filler words that contrasted /d/-/td/, /t/-/nt/, etc., whereas the Japanese listeners performed much poorer on the /t/-/nt/ contrasted words (70% accuracy without noise and 55% accuracy with a signal-to-noise ratio of -21 dB).

To assess the effect of training on noise tolerance, the participants who scored less than 70% accuracy for a noise-free /t/-/nt/ identification test were selected, and 15 days of training (two hours for each day) was administered without any noise added (Ueda et al., 2003). Again, the experiment is still going on. Current results are shown in Fig. 2. The effect of training was generalized into the noisy conditions. The interaction effects between signal-to-noise ratio and other variables were not significant.

**Beyond Auditory Scene Analysis**

What Bregman pointed out about speech perception in
infancy is important: an infant must work with auditory scene analysis, at least in its primitive form, to hear her mother's
voice out from the mixture of environmental sounds. Investigations of second language acquisition would contribute to
the study of auditory scene analysis in infancy, through elucidating the differences between the first and the second
language acquisition. Moreover, this line of investigation may clarify the difference between primitive segregation and
schema-based segregation.

Another point that is probably worth considering is the role of involuntary, or stimulus driven, attention. Although
Bregman considers primitive segregation, which is the core part of auditory scene analysis, is performed by a pre-attentive
process, this does not rule out the operation of involuntary attention. Especially when we think about the threading from
one perceptual modality to the other and its directionality (Driver and Spence, 1998), it is very hard not to use the
concept of involuntary attention. For example, one may notice an object, which was unattended and out of sight at that
moment, emitting a sound. The direction of the object is roughly estimated with the auditory information. Next, the
eyes and the head are moved toward the estimated direction. Finally, the eyes can focus on the object as the results of an
automatic process of auditory and visual scene analysis. The final stage includes a process that takes a proper correspondence between the auditory object and the visual object. The whole process can be triggered by the auditory sensation and finished by the visual sensation, but not vice versa (Driver and Spence, 1998).

Furthermore, it was confirmed that voluntary attention could control primitive segregation (Carlyon et al., 2001). Segregation builds up for some repeating triplet sequences: A stream consisted of both high and low tones is heard at the beginning, however, the stream segregates into two streams, a stream of only high tones and a stream of only low tones, after 10 seconds. This segregation started only after participants had directed their attention to repeating triplet sequences, when the sequences and noise were dichotically presented. Thus, primitive segregation is not entirely automatic.

One thing that is excluded—carefully perhaps—from the present framework of auditory scene analysis is the movement of
objects and ourselves. However, the present authors believe that, movement is also a building block of perceptual
organization. Information of movements of both objects and oneself can be organized as causality inference. The inference
may be based on a wide range of sensory inputs from auditory, visual, tactile, and proprioceptive sensation, and on internal
information in the brain. For example, if the cause is the movement of a talker in a certain direction, then the effect on a
listener side is the speech sound gradually shifting in that direction, so does every component of the speech sound at
the same time. Moreover, an auditory movement is almost synchronized with a visual movement. These covariations
could be utilized to form a more robust perceptual stream than a stream formed by auditory information alone. Sometimes,

auditory information modifies visual motion perception (Remijn et al., 2004).

Taking into account the cross-modal covariations caused by movements and both voluntary and involuntary attention may
be regarded as a natural extension of the ecological perspective (see also, Rosenblum, 2004). These two issues arguably play
some roles in our basic perceptual mechanisms, which are the very basis for our language perception and language
acquisition. Experimental investigations on these issues are further needed to clarify the mechanisms, from modality
specific, cross-modal, and modality neutral, i.e., cognitive or ecological (Rosenblum, 2004), point of views. Concerning
modality specificity, auditory scene analysis is a representative concept on the line of a traditional, modality specific point of
view, although it has a flavor of an ecological point of view.

Acknowledgements This work was supported in part by Grant-in-Aid for the 21st Century COE program, Grant-in-Aid
for Scientific Research No. 14101001 from the Japan Society for the Promotion of Science, and funding from the National
Institute of Information and Communications Technology.

References

Proc Inti-noise 86: 895–898
58: S35
631–637
Florentine M (1985) Speech perception in noise by fluent non-native listeners. Transactions of Technical Committee on
Psychological and Physiological Acoustics 11:55-16
Acoust Soc Jpn 54: 695–703 [In Japanese]
with controlled word predictability. J Acoust Soc Am 61: 1337–1351
Speech, Language, and Hearing Research 40: 686–693
Miller GA (1962) Decision units in the perception of speech. IRE Transactions on Information Theory IT-8: 81–83

Received: September 6, 2004
Accepted: November 2, 2004
Correspondence to: Kazuo Ueda, Department of Applied Information and Communication Sciences, Kyushu University, 4–9–1 Shiobaru, Minami-ku, Fukuoka 815–8540, Japan
Phone: +81–92–553–9460
Fax: +81–92–553–9460
e-mail: ueda@design.kyushu-u.ac.jp