ACOUSTICAL LETTER

Speech perception experiment using binaural integration of phonemic and prosodic information

Takayuki Arai[1,*], Chikashi Michimata[2] and Hirofumi Kamata[3]

[1]*Department of Information and Communication Sciences, Sophia University,*
7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan
[2]*Department of Psychology, Sophia University,*
7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan
[3]*Department of Psychology and Education, Wako University,*
2160 Kanai-cho, Machida, 195–8585 Japan

(*Received 20 May 2010, Accepted for publication 22 July 2010*)

Keywords: Binaural integration, Phonemic information, Prosodic information, Laterality of ears, Dichotic listening
PACS number: 43.71.An, 43.66.Pn [doi:10.1250/ast.32.34]

1. Introduction

Since Broadbent (1954) [1] first conducted a dichotic listening task by simultaneously presenting a pair of similar speech sounds at each ear, many aspects of dichotic listening have been studied, including Kimura's early interpretation from a neurological point of view [2]. In a dichotic listening task, a right-ear advantage is often observed, because language processing takes place in the left hemisphere of the brain, when the stimuli for the two ears are almost the same in terms of intensity, duration, and timing [3,4].

Zatorre also pointed out that the intonational information in a speech signal shows a right-hemispheric advantage [5,6]. In addition, Poeppel proposed the AST (asymmetric sampling in time) hypothesis of speech [7,8], where the time constant of temporal integration in the left-hemispheric auditory cortex is short (20–40 ms) and that of the right-hemispheric auditory cortex is long (150–250 ms).

In near-infrared spectroscopy (NIRS) studies [9,10], hemispheric laterality using pairs of stimuli contrasting intonational and phonemic information was discussed. The right-handed participants of this experiment showed stronger performance in the left auditory cortex when presented with phoneme-contrasting stimuli and the right auditory cortex when presented with intonation-contrasting stimuli.

On the basis of the studies of such laterality, the following evaluation method has been proposed to test central auditory processing disorders (CAPD) [11]:

- dichotic listening tests,
- monaural low redundancy speech tests,
- temporal patterning tests, and
- binaural interaction tests.

In binaural interaction tests, it is necessary to integrate the information coming from both ears, because a speech signal is segmented into uniform-length frames (e.g., 200 ms) and alternately presented to each of the ears by means of RASP (rapidly alternating speech perception) [12]. In the present study, we likewise conduct a binaural integration test in which phonemic and prosodic information of a speech signal are presented at each ear. In this test, a listener is only able to respond to a stimulus by integrating information from both ears. Because of the hemispheric laterality reported in the previous studies, the response time is expected to be shorter when phonemic information is presented at the right ear and prosodic information is presented at the left ear than the other way around.

2. Experiment

2.1. Speech samples

In this experiment, we used two types of Japanese speech samples: The declarative sentence "Kore wa, ___ desu ka. (This is ___.)" and the question "Kore wa ___ desu ka? (Is this ___?)." Because the former has a falling intonation and the latter has a rising intonation, we call them F and R, respectively. The target word "___" was a three-moraic word from Table 1. The words for the main session were selected from this table on the basis of the following criteria:

- the type of the accent is "no accent,"
- the vowels were the same within each pair of words, and
- the difference in word familiarity [13] between the two words within each pair was 0.5 or less.

For the training session, we prepared three extra pairs of words as follows: "manga-hanga," "tatami-katami," and "hakama-sakana."

We recorded a speaker pronouncing 48 sentences: 2 types of sentences × 24 pairs of words (18 pairs for the main session and 6 pairs for the training session). The speaker was a 22-year-old woman who had trained as an announcer. The recordings were made in a sound-treated room, and the speaker used a metronome to maintain a consistent speaking rate of 5 mora per second.

2.2. Stimuli

Each of the recorded speech samples (16 kHz sampling and 16 bit quantization) was processed to obtain two types of speech signals: The "ph signal" in which the prosodic information was suppressed, and the "pr signal" in which the phonemic information was suppressed. The ph signal is "noise-vocoded speech" based on linear predictive coding (LPC). We first extracted an LPC spectral envelope for each

[*] e-mail: arai@sophia.ac.jp

Table 1 The pairs of three-maraic words used in the main session of the perceptual experiment (values in parentheses indicate word familiarity).

tokei (5.8)	mokei (5.8)
kaseki (5.7)	zaseki (5.8)
kirin (6.0)	mirin (6.3)
unagi (5.8)	usagi (5.5)
tarako (6.4)	tabako (6.0)
kakashi (5.8)	karashi (5.3)
yanagi (5.9)	hayashi (6.1)
yubiwa (6.2)	kujira (5.8)
hamaki (5.3)	sanagi (5.3)

time frame, and a noise signal was used to excite the filter derived from the extracted envelope (the LPC order was 20). The pr signal is "hum speech" obtained using Praat software [14]. In this software, the glottal pulses are automatically estimated for voiced portions, and the impulse train is obtained as a source signal by putting the impulse function at each glottal pulse. Hum speech is obtained as output of a steady-state vocal-tract filter. It has formant frequencies similar to the "schwa" vowel after the source signal is fed into the filter. The root-mean-square values were normalized for all of the ph and pr signals.

Finally, we combined the ph and pr signals from the same speech sample to prepare a binaural stimulus. The <ph, pr> stimulus stands for a binaural signal where the ph signal is on the left and the pr signal is on the right. The <pr, ph> stimulus stands for a binaural signal where the pr signal is on the left and the ph signal is on the right. For each word, there are two types of sentences: F and R. For each sentence, we obtained the ph and pr signals. As a result, for each word, we have four stimuli: F<ph, pr>, F<pr, ph>, R<ph, pr>, and R<pr, ph>.

2.3. Participants

Nineteen listeners (13 males and 6 females), who are native speakers of the Tokyo dialect of Japanese and have little experience living abroad, participated in the experiment. They all have normal hearing and are right-handed, aged 20 to 22 years with an average age of 21.4 years. We determined that two participants (1 male and 1 female) out of the 19 were not able to understand the experimental procedure, and their data were excluded from the final results. Handedness was measured with the Edinburgh Handedness Inventory.

2.4. Procedure

The experiment was conducted in a sound-treated room, and a PC and headphones were used to present words and stimuli. The experiment was conducted in two sessions, each of which had a training session with 48 trials and a main session with 144 trials. In each trial, one binaural stimulus was presented through the headphones. First, we asked a participant to look at the "+" sign displayed at the center of the PC screen. As soon as the stimulus (either one of the four stimuli, i.e., F<ph, pr>, F<pr, ph>, R<ph, pr>, or R<pr, ph>, of a word within a word pair in Table 1) was presented, either one of the two words within the same word pair was displayed on the PC screen in Kana orthography. The participant was asked to judge two things within each trial: whether the sentence had an R (rising) or F (falling) intonation, and whether the target word of the stimulus presented through the headphones was the same as the one displayed on the PC screen.

With respect to the judgment of intonation, the participant was asked to answer only when the stimulus was R during the first session and F in the second session, or vice versa. The combination of the first/second session and R/F was counterbalanced among participants.

For the judgment of target word identification, the participant was asked to answer by hitting keys on a ten key pad. The four fingers, 1) left middle finger, 2) left pointer finger, 3) right pointer finger, and 4) right middle finger, were assigned to four keys lined up in the same row on the ten key pad. The participant was asked to simultaneously hit two keys with the left and right middle fingers (or pointer fingers) when the target word in the stimulus was the same as (or different from) the one displayed on the PC screen, or vice versa. The combination of middle/pointer fingers and same/different was counterbalanced among participants.

During the training session, the participant was notified as to whether the response was correct in each trial, and was also asked to become accustomed to the stimuli and the procedure. In the main session, we measured the reaction time (RT) as well as the correctness for each trial. We instructed the participants to attempt to make as correct and quick a response as possible. The measurement of the RT started when the word was displayed on the monitor screen. The system automatically moved to the next trial after two seconds if the participant did not respond.

2.5. Results and discussion

Because the correct rates showed a ceiling effect, we only looked at the RTs. The RTs for both the <pr, ph> and <ph, pr> stimuli were 670 ms on average; there was no significant difference due to ear laterality. However, as shown in Fig. 1, an interaction between the session and laterality was observed ($F(1,17) = 16.2$, $MSE = 19416.9$, $p < 0.001$). As seen in Fig. 1, the average RT did not change for <pr, ph> stimuli, whereas it was quicker in the second session compared with the first session for <ph, pr> stimuli.

3. Summary

To address the issue of ear laterality, we designed a binaural integration test, where a listener is able to respond to a stimulus only by integrating information from both ears. We expected the RT of the <pr, ph> stimuli to be shorter than the <ph, pr> stimuli. The experimental results of our study showed that the average RT in the first session followed the expectation; however, the average RT in the second session showed the opposite tendency. After a closer look, for <pr, ph> stimuli, the average RT did not show any difference between the first and second sessions. For <ph, pr> stimuli, on the other hand, the average RT became shorter as the session proceeded. It may be reasonable to interpret this as follows: the "advantageous" ear responds stably while the

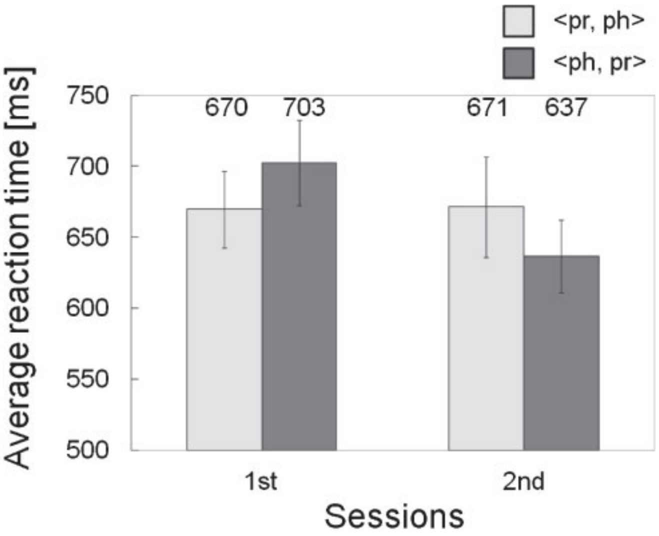

Fig. 1 Average reaction times to <pr, ph> and <ph, pr> stimuli in the first and second sessions.

"disadvantageous" ear is trained to process the presented information effectively. However, in a stimulus sentence, the timing at which a listener can judge the difference in phonemic information (the place where the target word is inserted) and the timing at which a listener can judge the difference in prosodic information (the sentence end) are different; therefore, the performance may have increased as a participant acquired a strategy for integrating the information presented to the two ears. In any case, further studies are needed before further discussion can be undertaken.

Acknowledgements

We would like to thank two alumni of the Arai Laboratory of Sophia University, Takuma Okuda and Keiko Kuwahara. We also thank Professor Hajime Tachiiri of the Ehime University for his useful comments. This research was partly supported by a Grant-in-Aid for Scientific Research (18530762) from the Japan Society for the Promotion of Science and by Sophia University Open Research Center from MEXT. This work was presented at the Spring Meeting of the Acoustical Society of Japan, in March 2009.

References

[1] D. Broadbent, "The role of auditory localization in attention and memory span," *J. Exp. Psychol.*, **47**, 191–196 (1954).

[2] D. Kimura, "Functional asymmetry of the brain in dichotic listening," *Cortex*, **3**, 163–168 (1967).

[3] R. B. Ivry and L. C. Robertson, *The Two Sides of Perception* (MIT Press, Cambridge, 1997).

[4] J. Ryalls, *A Basic Introduction to Speech Perception* (Singular Pub., San Diego, 1996).

[5] I. S. Johnsrude, R. J. Zatorre, B. A. Milner and C. Alan, "Left-hemisphere specialization for the processing of acoustic transients," *Neuroreport*, **8**, 1761–1765 (1997).

[6] R. J. Zatorre and B. Pascal, "Spectral and temporal processing in human auditory cortex," *Cerebral Cortex*, **11**, 946–953 (2001).

[7] D. Poeppel, "Pure word deafness and bilateral processing of the speech code," *Cognit. Sci.*, **25**, 679–693 (2001).

[8] D. Poeppel, "The analysis of speech in different temporal integration window: Cerebral lateralization as 'asymmetric sampling in time,'"*Speech Commun.*, **41**, 245–255 (2003).

[9] I. Furuya and K. Mori, "Cerebral lateralization in spoken language processing measured by multi-channel near-infrared spectroscopy (NIRS)," *No to Shinkei*, **55**, 226–231 (2003) (in Japanese).

[10] Y. Minagawa and K. Mori, "Near-infrared spectroscopic measurement of human language function," *Jpn. J. Clin. Psychiatry*, **33**, 741–747 (2004) (in Japanese).

[11] T. J. Bellis and J. M. Ferre, "Multidimensional approach to the differential diagnosis of central auditory processing disorders in children," *J. Am. Acad. Audiol.*, **10**, 319–328 (1999).

[12] J. A. Willeford and J. M. Bilger, "Auditory perception in children with learning disabilities," in *Handbook of Clinical Audiology*, 2nd ed., J. Katz, Ed. (Williams and Wilkins, Baltimore, 1978), pp. 410–425.

[13] S. Amano and T. Kondo, *Nihongo-no Goi-Tokusei* (*Lexical properties of Japanese*), (Sanseido, Tokyo, 1999) (in Japanese).

[14] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.*, **5**, 341–345 (2001).