A Multilingual Comparison of Intelligibility in Locally Time-Reversed Speech[1]

Kazuo UEDA*, Yoshitaka NAKAJIMA**, Wolfgang ELLERMEIER***, and Florian KATTNER***

*,**Dept. Human Sciences/Research Center for Applied Perceptual Science, Kyushu University, 4-9-1 Shiobaru Minami-ku, Fukuoka 815-8540, Japan

***Institut für Psychologie, Technische Universität Darmstadt, Alexanderstr. 10, D-64283 Darmstadt, Germany

E-mail: *ueda@design.kyushu-u.ac.jp **nakajima@design.kyushu-u.ac.jp

Abstract The purpose of the present investigation was to examine how the intelligibility of spoken sentences in Chinese, English, German, and Japanese changes as the duration of segments locally reversed in time increases. Thirty-five or eighteen sentences spoken by both a male and a female speaker in each language were extracted from a database (NTT-AT, Multilingual Speech Database 2002). The sentences were divided into segments of equal duration (20-170 ms); each segment was shaped with 7.5-ms cosine ramps, reversed in time, and subsequently joined with the other segments in the original order. The resulting utterances were presented diotically to 28 native speakers through headphones. The participants were instructed to write down what they heard without guessing. Percentage of correct syllables or morae was measured as an index of intelligibility. Intelligibility was above 90% when segment duration was 45 ms, but dropped with increasing segment duration to below 15% at 120 ms. The 50% intelligibility point was observed for segment durations of around 65-80 ms irrespective of language. This finding and the majority of the previous literature strongly suggest that a common time-averaging mechanism works in speech perception across different languages.

Keywords Spoken sentence, Segment duration, Syllable, Mora, Time-averaging mechanism

1 Introduction

A time-reversing technique has been applied in auditory experiments to produce stimuli that are unintelligible while the long-term spectra of them are maintained to be the same as in the originals (e.g., Jones, Miles, & Page, 1990; Howard & Poeppel, 2010). If a spoken sentence is successively segmented with a very short time interval, say 20 or 40 ms, and each segment is time-reversed, listeners can report the sentence perfectly; however, when the segment duration is lengthened to some extent, intelligibility declines. The technique, locally time-reversing, was first employed by Steffen and Werani (1994) to the best of our knowledge.

They aimed at determining the stimulus precision in time that is necessary in phoneme perception. They preserved the part below 300 Hz as the original, while they locally time-reversed the part above 300 Hz with 20- to 70-ms segments of an equal duration, because they constructed their stimuli in accordance with the framework of time perception proposed by Pöppel (1978): Phoneme perception should be determined within the time range of 20 to 100 milliseconds, while perception of syllables and intonation should be based on a longer time scale. A nonsense sentence in German, which was time-reversed with 70-ms segments, was presented first, then the segment duration was gradually shortened. They counted the number of participants who correctly wrote down the whole sentence at each time of stimulus presentation. The number of participants increased as the duration of a segment decreased. They concluded that the majority of their participants perceived the sentence perfectly less than 40 ms of segment duration. Al-

[1]Some parts of the work were presented at Fechner Day 2015, the 31st Annual Meeting of the International Society for Psychophysics in Québec, Canada, on 18 August 2015, at the 2nd Annual Meeting of the Society for Bioacoustics in Fukuoka, Japan, on 12 December 2015, and at the 49th Colloquium on Perception in Sendai, Japan, on 16 March 2016.

though they should be respected as pioneers of this line of investigations, their way of stimulus generation and intelligibility measurement probably made it difficult to interpret their results.

Saberi and Perrott (1999), seemingly being unaware of the work by Steffen and Werani, published a study about locally time-reversed speech as a brief communication in *Nature.* They simply time-reversed each 20- to 300-ms segment without filtering—a positive aspect of their method. Downsides of their method would be that what they asked from their seven participants was to rate the intelligibility of each stimulus, and that the stimuli were generated from a single sentence. Besides, it might be possible that the stimulus with the shortest segment duration—that should be perfectly understandable—was presented first, and then the other stimuli with longer segment duration were presented in an orderly fashion. Therefore, the results are likely to be biased.

A more objective measure of intelligibility, i.e., word intelligibility, was employed by the later investigators. Greenberg and Arai (2001, 2004) carefully avoided presenting a sentence in more than a single trial for each participant in their experiment of locally time-reversed speech in English employing 27 participants. Meunier, Cenier, Barkat, and Magrin-Chagnolleau (2002) employed 28 participants to make a comparison between the intelligibility curves of English and French—the English curve was provided by Greeenberg and Arai. Kiss, Cristescu, Fink, and Wittmann (2008), studying locally time-reversed speech in German, focused on performance difference between native and non-native listeners, 10 for each group.

Figure 1 summarized those previous data. The curve by Saberi and Perrott (1999) looks obviously odd. It is clear that those results obtained with the word intelligibility measure show rather similar tendency. Specifically, 50% of intelligibility was obtained at segment durations around 60-80 ms. Then, is it obvious that any other language shows a similar tendency like this?

It is common to categorize languages into three types of timing, that is, stress-, syllable-, and mora-timed languages. Ramus, Nespor, and Mehler (1999) proposed that those three types of timing could be identified by acoustical measurements (Fig. 2). The stress-timed languages, i.e., English, Polish, and Dutch, formed a cluster around the left side, the syllable-timed languages, i.e., Spanish, French, Italian, and Catalan, formed another cluster around the center, and the mora-timed language, i.e., Japanese, is located at the rightmost position in the graph, far away from any other languages. It would be possible that those differences in timing affect the intelligibility of locally time-reversed speech. For example, *obasan* (/o-ba-sa-N/,

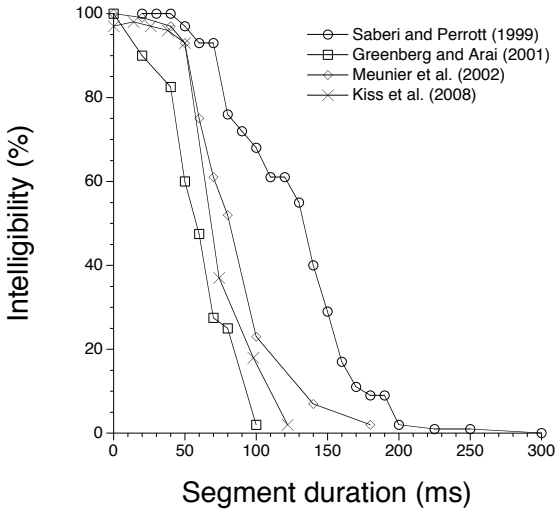

Figure 1 Summarized results of previous investigations about intelligibility of locally time-reversed speech. Only the data by 10 native listeners is presented here for Kiss et al. Note that the results by Saberi and Perrott (1999) were obtained with subjective ratings by seven participants.

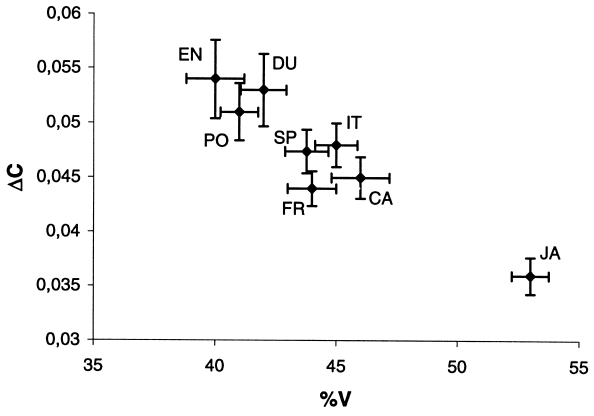

Figure 2 Three types of timing—stress, syllable, and mora—in various languages clustered with proportions of vocalic intervals in a sentence (horizontal axis) and the standard deviations of consonantal intervals (vertical axis). EN: English; PO: Polish; DU: Dutch; SP: Spanish; FR: French; IT: Italian; CA: Catalan; JA: Japanese. From Ramus et al. (1999).

a middle-aged lady or an aunt) has four morea, and *obāsan* (/o-ba-a-sa-N/, an old lady or an grandmother) has five morea in Japanese. The second vowel in both words is the same, i.e., /a/, however, the length of it in the latter word is doubled. The difference might sound subtle for non-native speakers (Hirata, 2004), but actually it makes an enormous difference in Japanese! This special distinction in Japanese may result in maintaining higher intelligibility than the other languages, because it is conceivable that vowels are less vulnerable than consonants when they are locally time-reversed.

Therefore, the purpose of the present investigation was to examine the intelligibility of spoken sentences in Chinese, English, German, and Japanese as the duration of locally time-reversed segments increases. We considered English and German as a kind of control, representing stress timed languages, Chinese as an example of a syllable-timed language, and Japanese as a mora-timed language. To make a fair comparison as far as possible, we measured syllable intelligibility rather than word intelligibility in Chinese, English, and German, and mora intelligibility in Japanese.

2 Method

2.1 Participants

Twenty-eight native speakers of Japanese, 10 females and 18 males (age, 21-23; median, 22 years), and 28 native speakers of Chinese, 17 females and 11 males (age, 21-32; median, 24 years), participated in the experiments performed in Kyushu, and 28 native speakers of German, 4 females and 24 males (age, 18-53; median 22 years), and 28 native speakers of English, 9 females and 19 males (age, 20-59; median 22 years), participated in the experiments performed in Darmstadt. All the participants passed a screening test with a clinical audiometer to ensure that they had normal hearing within the frequency range of 250-8000 Hz (International Organization for Standardization, 1998). They were divided into two groups, each of which was allotted to either the stimuli produced by a male speaker or a female speaker.

2.2 Stimuli

Thirty-five sentences in Chinese, English, German, and Japanese, spoken by a male and a female speaker, were extracted from NTT-AT Multilingual Speech Database 2002 (NTT-AT, Kawasaki, Japan; recorded with a 16-kHz sampling rate and 16-bit linear quantization) with eliminating unnecessary blanks and noises. Each sentences were segmented with a fixed interval, ranged from 20-170 ms including 7.5-ms cosine ramps. Then, each segment was reversed in time and subsequently joined in the original order.

Each sentence was used just in a single trial for each participant. Five sentences were allotted to each condition. The allotment of a block of five sentences was shifted among participants. The shift was continued until it reached the 14th participant, then the cycle was repeated from the beginning until the last, i.e., the 28th participant.

2.3 Apparatus

The stimuli were presented diotically through headphones [Beyerdynamic DT 990 (Beyerdynamic GmbH, Heilbronn, Germany)] in a soundproof booth [Kyushu: Music cabin SC3 (Takahashi Kensetsu, Kawasaki, Japan); Darmstadt: Industrial Acoustics Company (Niederkrüchten, Germany)]. An optical interface [Kyushu: USB interface Roland UA-4FX (Roland Corp., Shizuoka, Japan)] and a headphone amplifier with a built-in DA converter [Kyushu: Audiotechnica AT-DHA 3000 (Audiotechnica, Machida, Japan)], or a DA converter [Darmstadt: RME multiface II (Audio AG, Haimhausen, Germany)] and a headphone amplifier [Darmstadt: Behringer Pro 8 (Behringer, Zhongshan, China)] were used to drive the headphones. The sound pressure levels of the stimuli at the headphones were adjusted to 74 dB SPL with a 1-kHz calibration tone, which was provided with the speech database, by using an artificial ear [Brüel & Kjœr type 4153 (Brüel & Kjœr Sound & Vibration Measurement A/S, Nœrum, Denmark)], a condenser microphone (Brüel & Kjœr type 4192), and a sound level meter (Brüel & Kjœr type 2250).

2.4 Procedure

Each sentence was presented three times with an inter-stimulus-interval of 1 s on a trial before a participant started to answer. The participants were instructed to write down what they heard without guessing. The indices of intelligibility were a percentage of correctly reported syllables in Chinese, English, and German, and a percentage of correctly reported morae in Japanese.

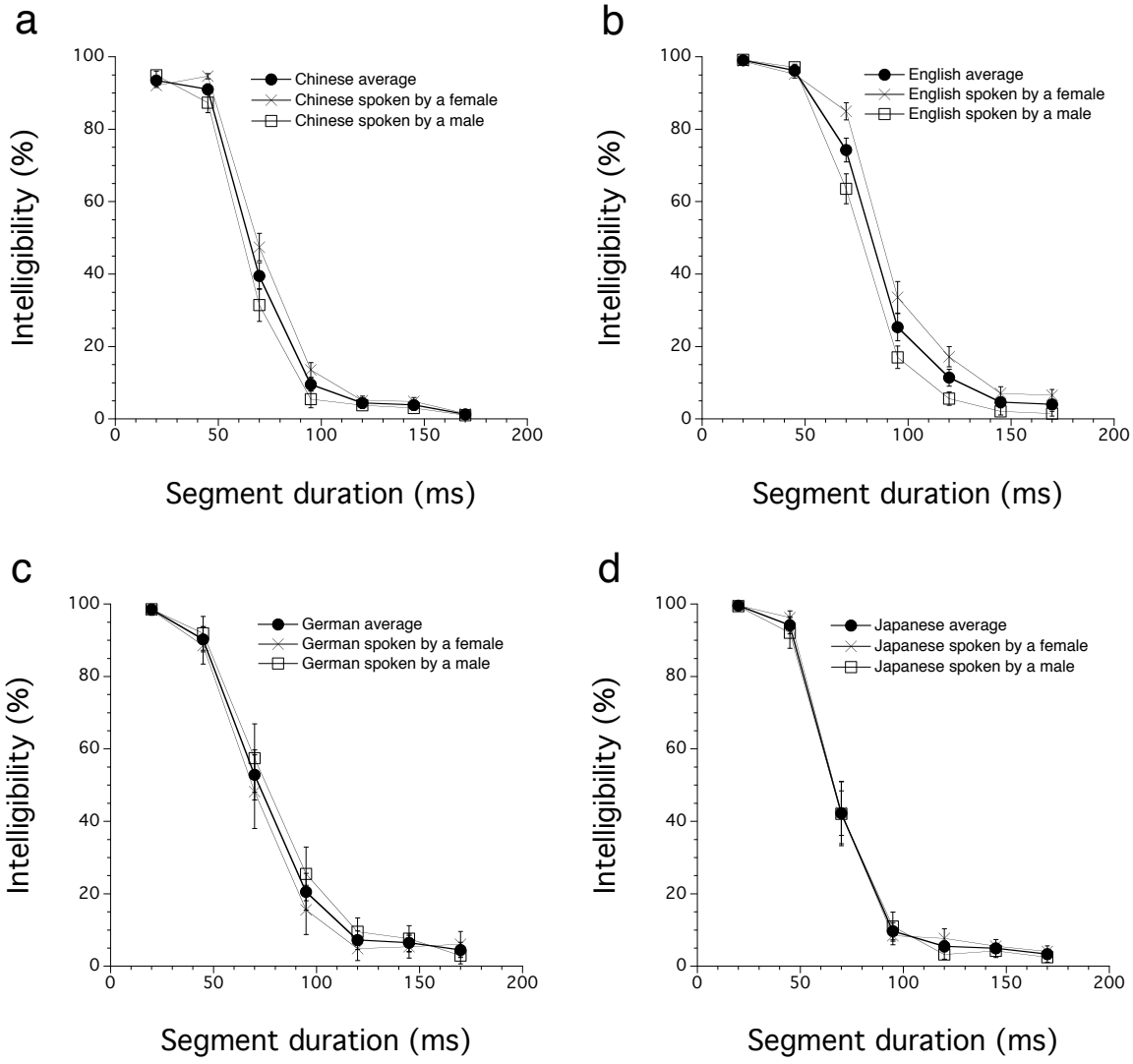

Figure 3 Percentage of syllable intelligibility in each language as a function of the segment duration obtained with 28 participants. All the data except in (**a**) were based on the test results with 35 sentences. **a,** Chinese results with 18 sentences. Those were selected from the data with 35 sentences because 17 unmatched sentences between the two talkers were mistakenly included in the experiment. **b,** English, **c,** German, and **d,** Japanese.

3 Results

Figure 3 shows the results. After finishing the experiments, we realized that 17 sentences out of 35 were unmatched between the two talkers in the Chinese experiment. The panel showing the Chinese results (Fig. 3a) therefore represents the data based on matched 18 sentences.

By and large, the curves showed similar tendency across different languages. The gender of the talkers did not show any consistent bias in the results.

General linear model (GLM) analysis revealed that the main effects of the segment duration were all significant [Chinese: $F(1, 495.9) = 868.80$, $p < 0.0001$; English: $F(1, 973.9) = 2954.99$, $p < 0.0001$; German: $F(1, 950) = 2059.42$, $p < 0.0001$; Japanese: $F(1, 950) = 2494.20$, $p < 0.0001$]. The effect of talkers was significant in English and German [English: $F(1, 47.75) = 43.50$, $p < 0.0001$; German: $F(1, 26) = 4.93$, $p = 0.035$], but not in Chinese and Japanese [Chinese: $F(1, 25.97) = 1.72$, $p = 0.20$; Japanese: $F(1, 26) = 2.23$, $p = 0.15$]. The interaction effects between the segment duration and the talkers were significant in English [$F(1, 973.9) = 3.93, p = 0.048$], but

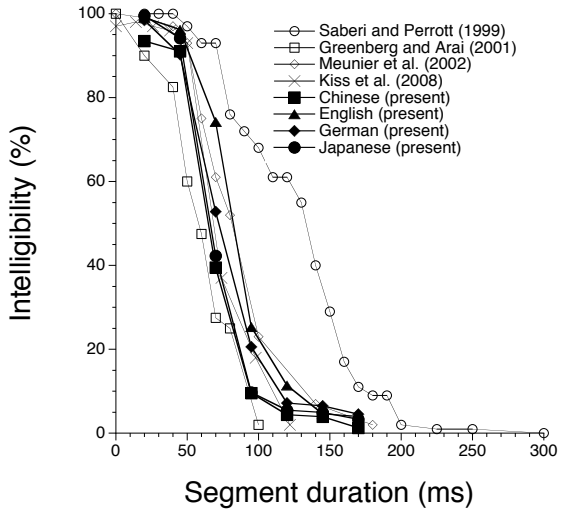

Figure 4 Summarized results of the previous and present investigations.

not in Chinese, German, and Japanese [Chinese: $F(1, 495.9) = 0.07$, $p = 0.80$; German: $F(1, 950) = 0.45$, $p = 0.50$; Japanese: $F(1, 950) = 0.47$, $p = 0.49$].

Three-way GLM analysis applied to the pooled data revealed that the main effects of language [$F(3, 101.4) = 52.38$, $p < 0.0001$], talker gender [$F(1, 112.4) = 6.98$, $p = 0.01$], and segment duration [$F(1, 3428) = 7136.99$, $p < 0.0001$] were all significant. The interaction effects of language × talker gender [$F(3, 104.5) = 8.40$, $p < 0.0001$] and language × segment duration [$F(3, 3374) = 35.01$, $p < 0.0001$] were significant, but talker gender × segment duration [$F(1, 3428) = 1.88$, $p = 0.17$] and language × talker gender × segment duration [$F(3, 3374) = 0.62$, $p = 0.60$] were not significant.

To seek a possible cause of the talker difference in English and German, duration differences of utterances were examined with paired t-tests (both sides). The analysis revealed that the female talker of English took significantly longer duration than the male talker [0.64 s longer than the male on average, $SE = 0.047$; $t(34) = 13.72, p < 0.0001$], and that the female talker in German took significantly shorter duration than the male talker [−0.06 s on average, $SE = 0.023$; $t(34) = -2.73, p = 0.01$]. No significant duration difference was observed in Chinese and Japanese [Chinese: −0.07 s on average, $SE = 0.058$; $t(17) = -1.24, p = 0.23$; Japanese: 0.09 s on average, $SE = 0.045$: $t(34) = 1.89, p = 0.07$].

Figure 4 shows the summary of the results available. The 50% intelligibility points in the present results fall in the range of 65-80 ms. Although the main effect of language was statistically significant, the curves of the present results and those of the previous results overlap each other, except Saberi and Perrott (1999). Moreover, the two English curves by Greenberg and Arai (2001, 2004) and the present investigation enclosed all the other curves in other languages except Saberi and Perrott.

4 Discussion

It is clear that speaking rates correlated with the small shifts of the curves between the talkers observed in Figure 3b and c: The talker who took longer duration–i.e., speaking more slowly–than the other made the curve shifted upward, and vice versa. It looks reasonable that a slower speaking rate resulted in a larger tolerance of intelligibility against lengthening segment duration.

Given that, the effects of language and talker gender might be actually caused by the difference in speaking rates, if it is possible to define a common measure of speaking rates across different languages in some way. The differences in timing of languages less likely affected the results.

Even if the differences of the intelligibility curves across the languages really exist, those curves show strikingly similar tendency. The results strongly suggest that a common time-averaging mechanism works in speech perception across different—stress-timed, syllable-timed, and mora-timed—languages. This mechanism is likely to involve lexical and semantic processing of speech, because Kiss et al. (2008) showed statistically significant differences between native and non-native listeners about the intelligibility of locally time-reversed speech.

Acknowledgments

The authors would like to thank Stephan Däbler, Ngar Nie Neo, Shunsuke Tamura, and Akihiko Shichida for their contribution in performing the experiment. This research was supported by Grants-in-Aid for Scientific Research Nos. 14101001, 19103003, 20330152, and 25242002 from the Japan Society for the Promotion of Science, and by a Grant-in-Aid for the 21st Century COE program from the Ministry of Education, Culture, Sports, Science and Technology.

References

Greenberg, S., & Arai, T. (2001). The relation between speech intelligibility and the complex modulation spectrum. In *Proceedings of the 7th European Conference on Speech communication and Technology (Eurospeech-2001)* (pp. 473–476).

Greenberg, S., & Arai, T. (2004). What are the essential cues for understanding spoken language? *IEICE Trans. Inf. & Syst.*, *E87-D*(5), 1059-1070.

Hirata, Y. (2004). Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts. *Journal of the Acoustical Society of America*, *116*(4), 2384-2394.

Howard, M. F., & Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *Journal of Neurophysiology*, *104*, 2500-2511.

International Organization for Standardization. (1998). *ISO 389-1. Acoustics—Reference Zero for the Calibration of Audiometric Equipment—Part 2: Reference Equivalent Threshold Sound Pressure Levels for Pure Tones and Supraaural Earphones.* International Organization for Standardization.

Jones, D. M., Miles, C., & Page, J. (1990). Disruption of proofreading by irrelevant speech: Effects of attention, arousal or memory? *Applied Cognitive Psychology*, *4*(2), 89–108.

Kiss, M., Cristescu, T., Fink, M., & Wittmann, M. (2008). Auditory language comprehension of temporally reversed speech signals in native and non-native speakers. *Acta Neurobiologiae Experimentalis*, *68*, 204-213.

Meunier, F., Cenier, T., Barkat, M., & Magrin-Chagnolleau, I. (2002). Mesure d'intelligibilité de segments de parole à l'envers en français. In *XXIVèmes Journées d'Étude sur la Parole, Nancy, 24-27 juin 2002.* Nancy.

Pöppel, E. (1978). Time perception. In R. Held, H. W. Leibowitz, & H.-L. Teuber (Eds.), *Perception* (pp. 713–729). Berlin, Heidelberg: Springer Berlin Heidelberg.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265-292.

Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature*, *398*(29 APRIL 1999), 760.

Steffen, A., & Werani, A. (1994). Ein Experiment zur Zeitverarbeitung bei der Sprachwahrnehmung. In G. Kegel, T. Arnhold, K. Dahlmeier, G. Schmid, & B. Tischer (Eds.), *Sprechwissenschaft & Psycholinguistik* (Vol. 6, pp. 189–205). Opladen: Westdeutscher Verlag.