French and English rhythms are perceptually discriminable with only intensity changes in low frequency regions of speech

Tsuyoshi Kuroda[†], Simon Grondin[†], Yoshitaka Nakajima[‡], Kazuo Ueda[‡]
[†]*Université Laval, Québec, Canada,* [‡]*Kyushu University, Fukuoka, Japan*
tsuyoshi.kuroda.1@ulaval.ca

Abstract

The purpose of this study was to determine which frequency band would contribute to discrimination between speech rhythms of French and English. Each trial consisted of two noises with different intensity changes. Each intensity change simulated the one that was derived from a frequency band of recorded sentences of French or English; the band had a center frequency of 350, 1000, 2150, or 4800 Hz. Participants evaluated the rhythm dissimilarity of two noises with an 8-point scale. Two noises were evaluated as more dissimilar when two sentences whose intensity changes were simulated by the noises were in different languages than when they were in the same language. Moreover, this tendency was reduced in 4800 Hz compared with the other bands. This indicates that French and English rhythms are discriminable with intensity changes of low frequency bands, even without any signs of pitch and phoneme.

Languages have their own rhythm. They are categorized according to what linguistic units comprise their rhythm. For instance, French is regarded as a *syllable-timed* language where syllables are regularly uttered while English is regarded as a *stress-timed* language where stressed syllables are regularly uttered (Fant, Kruckenberg, & Nord, 1991). However, the validity of this categorization is not well established by acoustic and psychological studies (see Patel, 2008). This seems due to the fact that there are few studies determining what acoustic property is linked to speech rhythm.

However, Nazzi, Bertoncini and Mehler (1998) demonstrated that newborn infants could discriminate different languages with only prosodic cues. In their experiment, speech stimuli were low-pass filtered to remove semantic cues but keep prosodic cues. Because newborn infants were naïve to semantic aspects of speech, their results indicated that languages could be discriminated even without semantic cues. Moreover, infants could discriminate between English (stress-timed) and Japanese (mora-timed) but not between English and Dutch (stress-timed). This supports the contemporary categorization of language rhythms and indicates that language rhythms are linked to some acoustic properties in low frequency regions of speech.

The method of Nazzi et al. (1998) could work only when infants were employed as participants, while the present study developed a method that could be used for adults to determine what acoustic property would be linked to speech rhythms of French and English. Note that Nazzi et al.'s stimuli included pitch cues because these included fundamental frequencies. However, the present experiment examined whether French and English rhythms could be discriminated with *only* temporal changes of intensity included in speech, and also examined which frequency band would be crucial for discriminating between French and English rhythms. In each trial, two noises were successively presented. These noises had different intensity changes. Each intensity change simulated the one that was derived from a frequency band of recorded sentences of French or English. Participants were

instructed to evaluate the rhythm dissimilarity of two noises. Note that the noises did not include any signs of pitch and phoneme of speech. If French and English rhythms are discriminated with only intensity changes included in speech, two noises should be perceived as dissimilar rhythms when one noise is derived from a French sentence and the other from an English sentence.

Method

Participants

Fifteen participants, five males and ten females aged 20-40 years, were recruited. They were students or employees at Université Laval. They consented to their participation by signing a form approved by the institutional ethical committee and received $100 CAN for their participation. Eight participants reported that they were French speakers and seven reported that they were English speakers.

Apparatus and stimuli

Recorded speech sentences which were sampled at 16000 Hz and quantized to 16 bits in an electric database (NTT-AT multilingual speech database 2002, 2002) were used. Four French and four British English sentences were randomly selected from the database for each participant; however, only sentences whose duration was approximately between 1.6 and 2.8 s were selected. Since each sentence was spoken by five males and five females in each language in the database, two males were randomly selected for two of the four selected sentences and two females were randomly selected for the other sentences, i.e., sentence 1 was spoken by male 1, sentence 2 by male 2, sentence 3 by female 1, and sentence 4 by female 2.

The procedure of making stimuli is illustrated in Figure 1. Each sentence passed through a band-pass filter, whose parameter was decided with Bark scale (Scharf & Buus, 1986). The filter always had a range of 2 Bark while its center was located on 4, 9, 14, or 19 Bark. The filter condition is called with a frequency corresponding to a center bark of the filter, i.e., 350, 1000, 2150, and 4800 Hz.

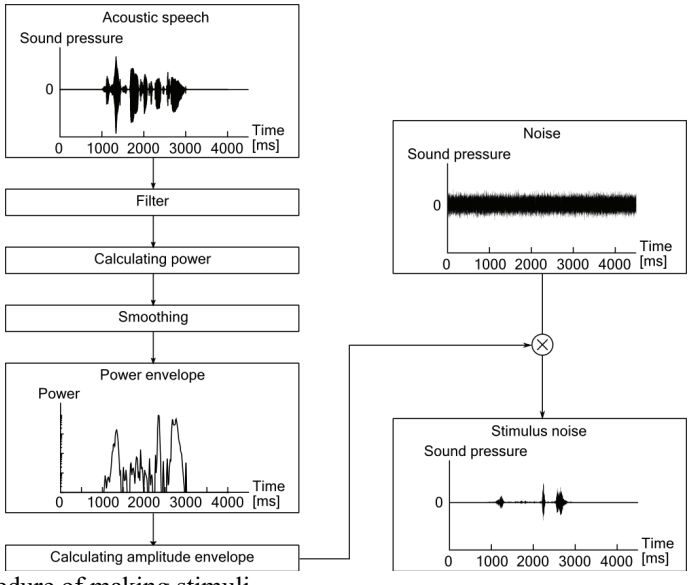

Figure 1. Procedure of making stimuli.

Amplitude (sound pressure) at each sample of the filtered speech was squared to obtain a power envelope. The resulting envelope was smoothed by calculating a moving average with a Gaussian window whose standard deviation was 5 ms. Then, a white noise of 150-7000 Hz was generated and its amplitude was multiplied by a square root of power at each sample of the smoothed envelope; consequently, the noise had the same power envelope as the filtered speech. There were 32 noises (2 languages × 4 sentences × 4 filters) in total.

Stimulus intensity was calibrated in two conditions. In one, the root-mean-square (RMS) level of each noise was calibrated at 25 dB sensation level (SL), i.e., all noises had equal RMS levels (mean-equal condition). In the other, the peak intensity of each noise was calibrated at 35 dB SL, i.e., all noises had equal peak levels (peak-equal condition). The reason why these conditions had different reference levels (25 and 35 dB SL) is because the mean-equal condition makes stimuli even louder than the peak-equal condition if these conditions have equal reference levels. Different observers were allocated to each of the two conditions, i.e., these conditions were between-participants conditions. Four French-speaking and three English-speaking participants were allocated to the mean-equal condition while four French-speaking and four English-speaking participants were allocated to the peak-equal condition.

Procedure

Two noises were successively presented in each trial. Participants were instructed to evaluate the rhythm dissimilarity of two noises with an 8-point scale, where "1" indicates "exactly the same" and "8" indicates "extremely dissimilar." Scales of around 8 points were used in multidimensional-scaling studies (e.g., Abelson, 1954-55). Participants responded by clicking on a pane on a computer display. They listened to stimuli by clicking on the "play" pane. They were allowed to listen to stimuli only once in each trial, but when listening was disturbed for some specific reason (e.g., yawning or coughing), they could listen to the stimuli again by clicking on the "replay" pane. Two noises were separated by an inter-stimulus interval varied from 2.5 to 3 s randomly.

Because the same noise was not doubly presented within one trial, there were 992 trials (${}_{32}P_2$). The experiment took two sessions each consisting of the 992 trials, i.e., participants responded twice for each of the 992 trials. The order of the trials was randomized in each session. Each session was divided into 16 blocks each consisting of 62 trials. Before the beginning of the experiment, threshold intensity for detecting a noise was measured and a practice block consisting of 62 randomly selected trials was carried out. Participants completed the threshold measurement and the practice block in one day while they completed the experimental sessions over 16 days (two blocks per day). Thus, the experiment was completed over 17 days. Each block took about 20 minutes, and a break of a few minutes was taken between blocks in each day. There were two warm-up trials at the beginning of each block and these trials consisted of the stimuli that were to be presented in the last two trials of the block.

Results

Each participant made four responses for each of the 496 noise pairs; the order of two successive noises was collapsed (992 trials ÷ 2 orders). These four responses were averaged. Kruskal's nonparametric multidimensional scaling was conducted on a dissimilarity (triangular) matrix for each participant (with software of R version 2.14.1). The reason why this analysis was conducted on individual data instead of pooled data is because four sentences were randomly selected for each participant, i.e., participants had different sets of sentences.

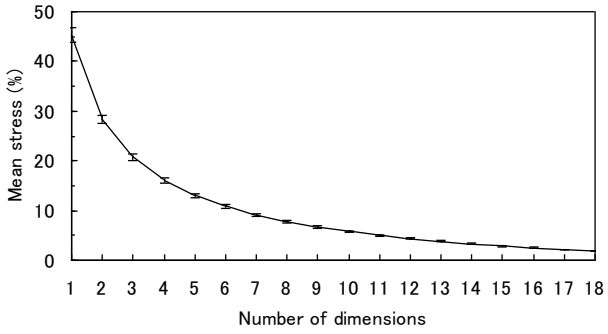

Figure 2. Mean stress in each number of dimensions that were constructed by a multidimensional scaling. Bars are standard error of mean.

The scaling approach has been taken in literatures to visualize relations between stimuli in a space of a few dimensions. Unfortunately, in the present experiment, the *stress* was too high, i.e., the goodness of fit was too poor, to visualize the data with a few dimensions (Figure 2). As an alternative method for approaching the data, the scaling was constructed with as many dimensions as possible; it was constructed with eighteen dimensions where the *stress* was sufficiently low (1.9%). In this approach, the multidimensional scaling was utilized simply for converting the nonparametric dependent variables, i.e., dissimilarity ratings, into the parametric ones; the number of dimensions had to be increased to minimize the stress.

If French and English rhythms are discriminated with power fluctuations in a limited frequency region, a Euclidian distance between French and English sentences, which were located in the 18-dimentional space, should be increased in a specific band relative to the other band conditions. Because there were four English and four French sentences, we calculated 16 Euclidian distances for 16 pairs of English vs. French sentences (4 × 4) in each band condition. These 16 distances were squared and summed up, the resulting value indicating how far apart English and French conditions were in each band in the 18-dimensional space (the *two-language separation* — TS). In addition, 28 Euclidian distances for all 28 pairs of sentences in each band (${}_8P_2$ because there were 8 sentences, i.e, 4 French + 4 English sentences) were squared and summed up, the resulting value indicating how large a divergence between individual sentences was (the *individual-sentence divergence* — ID). Dividing TS by ID gave an index of *relative separation* (RS). Note that RS becomes around .57 if all pairs of sentences lead to equal distances; for example, if all pairs lead to a distance of 1, TS becomes 16 and ID becomes 28, resulting in RS of around .57 (16 ÷ 28). In other words, RS above .57 indicates that two power envelopes (noises) were evaluated as more dissimilar rhythms when these envelopes were obtained from sentences in different languages than when obtained from sentences in the same language (see Nakajima & Takeichi, 2011, for the similar approach with correlation matrix).

Mean RSs in each experimental condition are shown in Figure 3. Note that in this figure “French” and “English” means participants’ speaking language instead of the four sentences’ language. In general, the 350-, 1000- and 2700-Hz conditions led to RSs higher than .57 except French-speaking participants in the peak-equal condition. A *t* test was conducted to examine whether each band led to a significantly higher RS than .57, i.e., with a null hypothesis that the mean RS for each band was .57. Because the between-participants conditions were pooled, the test was conducted four times for the four bands. A significant

difference was obtained in 350 Hz [$t(14) = 2.723$, $p = .016$], 1000 Hz [$t(14) = 3.252$, $p = .006$], and 2150 Hz [$t(14) = 2.931$, $p = .011$], but not in 4800 Hz [$t(14) = -.570$, $p = .578$].

An ANOVA according to 2 (participants' language) × 2 (power) × 4 (band) design, with repeated measures on the last factor, revealed that the band effect, $F(3, 33) = 4.708$, $p = .008$, $\eta_p^2 = .300$, as well as the power effect, $F(1, 11) = 8.066$, $p = .016$, $\eta_p^2 = .423$, was significant, while the language effect was not significant, $F(1, 11) = 2.027$, $p = .182$, $\eta_p^2 = .156$. The interaction between the power and the band was significant, $F(3, 33) = 3.023$, $p = .043$, $\eta_p^2 = .216$. The interaction between the language and the power was marginally significant, $F(1, 11) = 3.744$, $p = .079$, $\eta_p^2 = .254$. The remaining interactions were not significant ($F < 1.6$).

Because the interaction between the power and the band was significant, simple main effects were examined. The mean RS changed significantly depending on the frequency band in the mean-equal condition, $F(3, 33) = 7.004$, $p < .001$, $\eta_p^2 = .389$, but not in the peak-equal condition, $F(3, 33) = .728$, $p = .543$, $\eta_p^2 = .062$. In addition, the mean-equal condition led to a significantly higher RS than the peak-equal condition in 350 Hz, $F(1, 44) = 11.181$, $p = .002$, $\eta_p^2 = .203$, and in 1000 Hz, $F(1, 44) = 4.132$, $p = .048$, $\eta_p^2 = .086$.

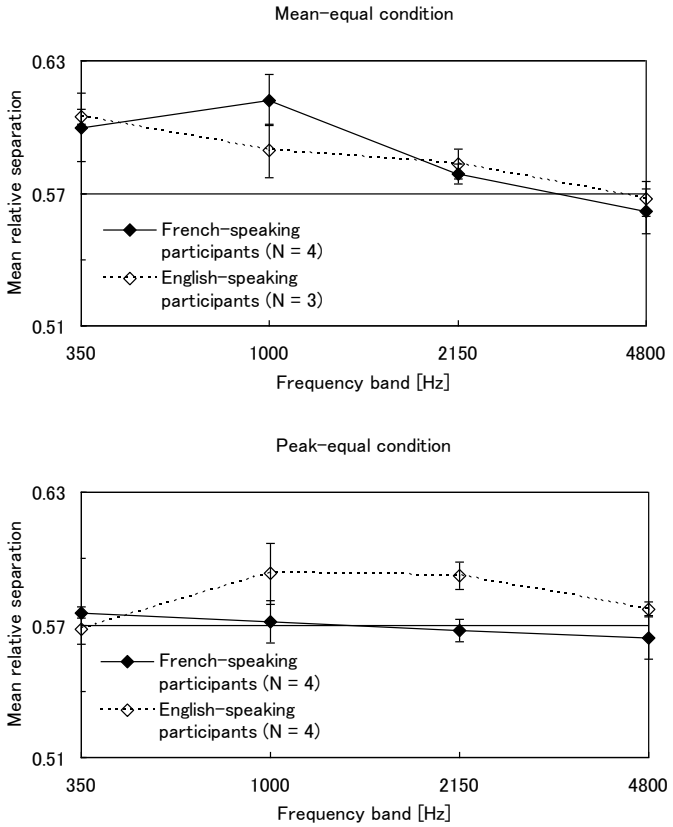

Figure 3. Mean relative separations in each frequency band for the mean-equal condition (upper panel) and for the peak-equal condition (lower panel). Bars are standard error of mean. Note that "French" and "English" means participants' speaking language, instead of the four sentences' language.

Discussion

The present study examined whether French and English rhythms could be discriminated with only temporal changes of intensity included in speech. To avoid presenting any other perceptual cues, e.g., phoneme and pitch, included in speech, the present experiment employed noises that had the same power envelopes as each frequency band of speech had. Because this approach focused on factors determining speech rhythms of French and English, it did not aimed to explain the whole aspects of perceptual difference between these languages. However, the results of the present study indicated that, even without any signs of phoneme and pitch, participants could perceive two power envelopes as more dissimilar rhythms when these envelopes were obtained from sentences in different languages than when obtained from sentences in the same language.

Moreover, the tendency to perceive French and English envelopes as dissimilar was reduced in 4800 Hz compared with the other bands, especially when these envelopes were calibrated at equal RMS levels (in the mean-equal condition). This indicates that French and English rhythms are determined by intensity changes in frequency regions below 2150 Hz. Since this frequency region determines the temporal arrangements of language nucleus, i.e., speech rhythms, of French and English, intensity below 2150 Hz could be utilized for constructing a physical parameter expressing *sonority*, which determines the temporal frames of syllables in phonology. This proposition is consistent with that proposed by Nakajima, Ueda, Fujimaru, Motomura and Ohsaka (2012) who investigated acoustical correlates of sonority in British English with factor analytical approaches.

Acknowledgements

This research was made possible by a research grant awarded to SG by the Natural Sciences and Engineering Council of Canada and grants awarded to YN and to KU by Japan Society of the Promotion of Science.

References

Abelson, R. P. (1954-55). A technique and a model for multi-dimensional attitude scaling. *Public Opinion Quarterly*, *18*, 405-418.

Fant, G., Kruckenberg, A., & Nord, L. (1991). Durational correlates of stress in Swedish, French, and English. *Journal of Phonetics*, *19*, 351-365.

Nakajima, Y., & Takeichi, H. (2011). Human processing of short temporal intervals as revealed by an ERP waveform analysis. *Frontiers in Integrative Neuroscience*, *5*, 2-10.

Nakajima, Y., Ueda, K., Fujimaru, S., Motomura, H., & Ohsaka, Y. (2012). Acoustic correlate of phonological sonority in British English. *Proceedings of the 28th Annual Meeting of the International Society of Psychophysics*. in preparation.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward and understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 756-766.

NTT-AT multilingual speech database 2002 [CD-ROM]. (2002). Tokyo: NTT Advanced Technology, Co.

Patel, A. D. (2008). *Music, Language, and the Brain*. New York: Oxford University Press.

Scharf, B., & Buus, S. (1986). Audition I: Stimulus, physiology, thresholds. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (pp. 14-1 to 14-71). New York: John Wiley & Sons.