LANGUAGE DISCRIMINATION USING LOW-PASS FILTERED SONGS: PERCEPTION OF DIFFERENT RHYTHM CLASSES

Hajime Takeyasu & Noriko Hattori

Mie University, Japan
thonsei@yahoo.co.jp; hattori@human.mie-u.ac.jp

ABSTRACT

As a characterization of language rhythm typology, several indices have been proposed. Among them are nPVI, rPVI, %V, ΔV, and ΔC. Investigations of their usefulness as indicators of language rhythm classes are generally based on production data. In this paper, a perceptual experiment is reported using low-pass filtered songs sung in English and Japanese. The aim was to examine which index best reflects the discrimination between the two different rhythm classes under the conditions in which segmental and pitch cues are reduced while rhythmic properties are preserved. The result showed that %V is the most reliable index for the discrimination of English and Japanese.

Keywords: rhythm, perception, low-pass filtered songs, %V, nPVI

1. INTRODUCTION

The languages of the world are categorized into a small number of rhythm classes: a traditional dichotomy between stress-timed and syllable-timed languages with the later addition of mora-timed languages as a third rhythm class [1, 5, 7]. Several measurements have been proposed which claim to be acoustic correlates of linguistic rhythm. The major indices suggested so far are as follows [2, 6, 9, 12, 13, 14, 15, 16, 18]:

- nPVI (the normalized pairwise variability index): the degree of contrast between successive durations in an utterance
- rPVI (raw pairwise variability index): computed the same way as the nPVI but without normalization
- %V: the proportion of vocalic intervals in the sentence (vocalic interval = section of speech between vowel onset and vowel offset)
- ΔV: the standard deviation of vocalic intervals
- ΔC: the standard deviation of consonantal intervals

Perceptual studies [11, 16] have shown that newborn babies as well as adults can discriminate between languages belonging to different rhythm classes. The rationale behind these cross-linguistic perceptual experiments which used filtered speech is that if the listener can tell two languages apart when the only cues are rhythmic, then the languages belong to distinct rhythmic classes [17]. Conversely if two languages have different types of rhythm, native listeners should be able to discriminate these two languages, on the basis of the rhythmic cues [17]. This paper investigates whether this also applies to sung speech.

In this study, English and Japanese were chosen as prototypical examples of stress-timed and mora-timed languages respectively. A perceptual experiment using low-pass filtered songs was carried out to test our hypothesis that adult native listeners could discriminate the two prototypical languages in terms of rhythm on the basis of rhythmic cues in songs as well as spoken language.

2. EXPERIMENT

The current study using low-pass filtered songs was designed to test to what extent the indices used to explain the production data would apply to the perceptual data.

2.1. Method

2.1.1. Subjects

Twenty-five adult native speakers of Japanese with normal hearing participated in this experiment.

2.1.2. Stimuli

All stimuli were created from the song "Teru-no uta" (from a Japanese film, *Gedo Senki* 'Tales of Earthsea' directed by Goro Miyazaki) sung in either Japanese or English (dubbed), which is available as a DVD. We selected this song because it is unaccompanied and its voices are easy to trace for acoustic measurements. The songs were low-pass filtered with a cutoff frequency of 500Hz to reduce segmental information while preserving the melody. Then, both the English and the Japanese songs were divided into 20 parts with 2 bars each,

resulting in 20 stimuli with 2 bars each for each language. Each pair of stimuli has almost the same melodies. They cannot be exactly the same because different lyrics in English and Japanese sometimes require different number of musical notes. The number of syllables in each stimulus is 11 or 12 for English, and 10 to 12 for Japanese. The discrepancy in the number of syllables between English and Japanese lyrics occurred when the latter had long vowel and/or geminate consonants (This has nothing to do with the results of the experiment. See Section 2.2.3.).

A total of 40 stimuli was thus prepared: 20 stimuli originally sung in English (hereafter 'English-series') and 20 in Japanese (hereafter 'Japanese-series'). Prior to the perceptual experiment, the durations of each segment were measured for each stimulus to calculate rhythmic indices. The current study followed the segmentation method presented in [15]. That is, segmentation of the songs was based on a display showing both the waveform and spectrogram on the Praat [3], plus interactive playback. The results of the acoustic analyses will be discussed in Section 2.2.1.

2.1.3. Procedure

The participants were tested individually in a quiet room. Each stimulus was presented to listeners once in random order and listeners were asked to decide whether the stimulus was originally sung in Japanese or English.

2.1.4. Predictions

Language rhythm is closely related to its syllable structures [5]. Compared with Japanese, English allows by far more complex syllable structures at the onset and the coda. It can be predicted from the fact that the ratio of durations occupied by consonants in a given length of an utterance is bigger in English (that is, English-series will have lower values of %V than Japanese) if these differences in syllable structures are also maintained in sung speech. Furthermore, English has a wider variety of syllable structures than Japanese, which necessarily leads to a greater variability in adjacent syllable durations (that is, English-series will have greater values of indices such as ΔV, ΔC, nPVI, and rPVI). Also, the differences in their values are expected to be used as cues in perception. These predictions will be tested against the results.

2.2. Results

2.2.1. Acoustic analyses

To examine whether our English- and Japanese-series are acoustically different (in terms of the rhythmic indices), we measured the following factors referred to in Section 1 (and, additionally, some other factors which might be related to the discrimination of rhythm classes) for each stimulus: %V, nPVI of V, ΔV, mean duration of V (Mean V), coefficient of V variation, ΔC, rPVI of C, nPVI of C (the formulae for these indices, see [13]). For each rhythmic index, paired *t*-tests were performed. The results of measurements are given in Table 1.

Table 1: Results of acoustic analyses of the stimuli.

	English-series	Japanese-series	t-test (English vs. Japanese)		
	Mean (SD)	*Mean (SD)*	*ES (d_z)*	*t(19)*	*p (two-tailed)*
%V	.66 (.02)	.80 (.04)	**-3.720**	**-16.290**	**.000**
nPVI of V	43.06 (14.79)	34.30 (12.90)	**.554**	**2.471**	**.023**
ΔV	156.03 (35.60)	186.81 (64.16)	-.360	-1.642	.117
Mean V	339.42 (16.79)	420.78 (52.20)	**-1.842**	**-7.441**	**.000**
Coeff. of V variation	.46 (.11)	.44 (.10)	.154	.689	.499
ΔC	79.67 (13.05)	56.95 (31.51)	**.900**	**3.612**	**.002**
rPVI of C	9638.21 (1884.31)	6622.50 (3833.00)	**.784**	**3.424**	**.003**
nPVI of C	58.00 (11.33)	54.54 (19.26)	.142	.641	.529

Five out of eight mean rhythmic indices turned out to be significantly different for English-series and Japanese-series. When the differences in English- and Japanese-series are converted into an effect size (*ES*) measure for paired *t*-test, $d_{z'}$ [4], ES value of %V was the highest ($d_{z'}$ = -3.720), followed by Mean V ($d_{z'}$ = -1.842), ΔC ($d_{z'}$ = .900), rPVI of C ($d_{z'}$ = .784), and nPVI of V ($d_{z'}$ = .554), which suggests that %V is the most reliable index for the discrimination of English/Japanese.

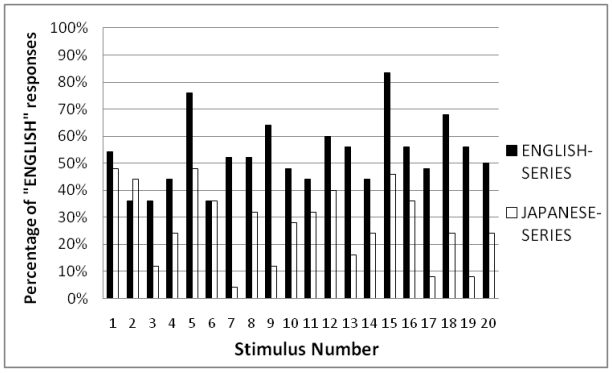
No results conflicted with our predictions in Section 2.1.4. This suggests that the rhythmic characteristics of English and Japanese due to the differences in the permitted syllable structures are maintained in our sung stimuli, and these rhythmic indices can be effective cues to the discrimination of the two languages even in songs.

2.2.2. Listeners' responses

Figure 1 shows percentages of the "English" responses for each stimulus. As shown in Figure 1, when compared to the corresponding stimuli of Japanese-series, most of the stimuli of English-series were more likely to be judged as "English". The mean percentage of "English" responses was 53.1% for the English-series and 27.3% for the Japanese series, and a paired t-test revealed that there was a significant difference between the two

sample means [*t(19)* = -7.040, $p < .001$, lower 95% CI limit = -0.336, upper 95% CI limit = .182, *ES* = 1.934]. Since the stimuli retained little segmental information, this result suggests that native listeners of Japanese responded according to suprasegmental information such as %V, nPVI, etc.

Figure 1: Listeners' responses for each stimulus.



2.2.3. Correlation analyses and regression analyses

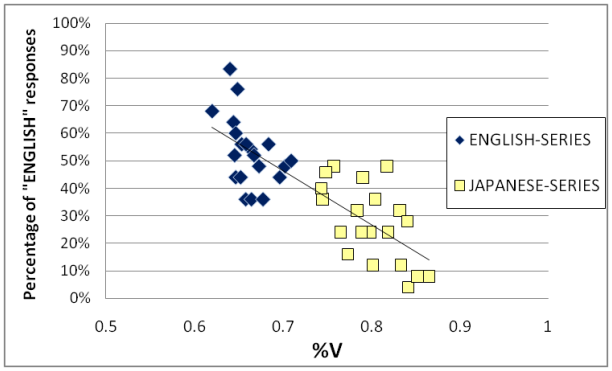
In order to determine which type of information in the stimuli was most responsible for the listeners' judgments, three types of correlation coefficients (Pearson's product-moment correlation coefficient (*r*), Spearman's rank-order correlation coefficient (r_s), and Kendall's tau coefficient) were calculated. Because the three types of correlation analyses resulted in the same conclusion, only Pearson's correlation coefficients are shown in Table 2.

Table 2: Correlation Coefficients between the percentage of "English" responses and rhythm factors.

	%V	**nPVI of V**	ΔV	**Mean V**	Coeff. of V variation	**ΔC**	**rPVI of C**	nPVI of C
Pearson's r	**-.802**	**.412**	-.251	**-.676**	.149	**.477**	**.562**	.309
p (two-tailed)	**.000**	**.008**	.118	**.000**	.359	**.002**	**.000**	.052
N	**40**	**40**	40	**40**	40	**40**	**40**	40

A series of correlation analyses revealed that five out of eight factors were significantly correlated with the percentage of "English" responses. Among the five significant factors, the Pearson's correlation coefficient (in terms of absolute value) of %V was the highest ($r = -.802$), followed by Mean V ($r = -.676$), rPVI of C ($r = .562$), ΔC ($r = .477$), and nPVI of V ($r = .412$), which suggests that %V is the strongest factor related to the listeners' judgments. Figure 2 provides an example of a scatter plot for %V and listeners' responses. The rank order parallels to the one in our acoustic analyses (see Section 2.2.1) except for the order of rPVI of C and ΔC, suggesting that rhythmic indices built on acoustic measurements are strongly related to perception of rhythm.

Figure 2: Scatter plot of %V and listeners' responses.



The fact that five factors are significantly correlated with listeners' responses, however, does not necessarily mean that we need all the five factors to describe our results (listeners' response patterns) because some of the factors discussed in this paper are highly correlated and may provide redundant information. Considering multi-collinearity, to find the best model explaining our results, we performed a stepwise logistic regression analysis with the five factors above as independent variables and listeners' responses (English or Japanese) as a dependent variable.

Table 3 shows the result of the stepwise logistic regression analysis. It turned out that only %V ($B = -7.606$, $W^2 = 50.434$, $df = 1$, $p < .001$) and rPVI of C ($B = 5.144 \times 10^{-5}$, $W^2 = 4.675$, $df = 1$, $p < .05$) were substantially significant. Change in the value of *-2 log likelihood ratio* when the factor was excluded from the model was 52.565 for %V and 4.578 for rPVI of C, which suggests that %V has stronger effect on listeners' responses than rPVI has although both are statistically significant. The other factors were excluded from the model possibly due to multicollinearity with either %V or rPVI of C.

Table 3: Selected model and estimated parameters.

	B	*SE*	*Wald*	*df*	*p (two-tailed)*	*Exp(B)*
%V	-7.606	1.071	50.434	1	.000	.000
rPVI of C	5.144×10^{-5}	.000	4.675	1	.031	1.000
Intercept	4.696	.886	28.104	1	.000	109.513

The same model was obtained when we further add the number of the musical notes and/or syllables in each stimulus as independent variables in the regression model. This suggests that although there were some differences in the melody between corresponding members of the stimulus-series (see

Section 2.1.2), the results reported here cannot be attributed to them in any great measure.

3. DISCUSSION

The English and Japanese stimuli used in this perceptual experiment are those with reduced segmental cues while preserving suprasegmental information. The counterpart stimuli in either language share almost the same melodies, since the arrangements of musical notes in their original songs are the same. This means that listeners cannot rely on their pitch contours in judging the stimuli. Thus the current experiment suggests that the listeners categorize the stimuli into English and Japanese based on suprasegmental information other than pitch cues. Rhythmic information should be a likely candidate. As a strong factor related to the listeners' judgments, %V and rPVI of C were found to be the most reliable indices among others. The result shows the status of %V as the strongest index in the discrimination of the two languages.

This finding in terms of perception matches the results from the production data. The indices proved to be reliable in the production data are also valid for the perception data. The fact that %V and rPVI of C were important in our experiment can be attributed to the differences in the permitted syllable structures in the two languages. The results of our experiment suggest that these differences in syllable structures are reflected even in songs and that phonological units such as syllable and its structure are essential in discussing language rhythm. According to Maddieson's 'Syllable Complexity Scale' [10], which evaluates the relative complexity of the maximal permitted structure of onsets, nuclei and codas separately, English is at the top of the scale with 8 points (most complex), while Japanese is at level 4 (less complex). In the light of its importance as a contributing factor to rhythm typology [8], the differing complexity of syllable structures in English and Japanese should generate the differences in the rhythmic indices, which in turn affected our listeners' judgments. Thus, the finding presented here is in line with the idea that the syllable structure should affect speech rhythm.

4. CONCLUSIONS

The experiment reported here has demonstrated the reliability of %V and rPVI of C as indices in the discrimination of English and Japanese presented with limited segmental cues. Further studies will be necessary to see if their reliability as indices also applies to filtered spontaneous speech. It will also be worth investigating whether the rank order of these indices remains the same or varies depending on the language pairs to be used in the discrimination experiment.

5. ACKNOWLEDGEMENTS

This work is supported by a Grant-in-Aid for Scientific Research (C) (Japan Society for the Promotion of Science, Grant No. 21520506) to the second author.

6. REFERENCES

[1] Abercrombie, D. 1965. *Studies in Phonetics and Linguistics*. London: Oxford University Press.

[2] Barry, W., Andreeva, B., Koreman, J. 2009. Do rhythm measures reflect perceived rhythm? *Phonetica* 66, 78-94.

[3] Boerma, P. Praat: Doing phonetics by computer. *http://www.fon.hum.uva.nl/praat/*

[4] Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Psychology Press.

[5] Dauer, R.M. 1983. Stress-timing and syllable-timing reanalysed. *Journal of Phonetics* 11, 51-62.

[6] Grabe, E., Low, E.L. 2002. Durational variability in speech and the rhythm class hypothesis. In Gussenhoven, C., Warner, N. (eds.), *Laboratory Phonology* 7. Berlin: Mouton de Gruyter, 515-546.

[7] Jones, D. 1972. *An Outline of English Phonetics*. Cambridge: Cambridge University Press.

[8] Ladefoged, P. 2006. *A Course in Phonetics* (5th ed.). Boston: Thomson.

[9] Low E.L., Grabe, E., Nolan, F. 2000. Quantitative characterization of speech rhythm: Syllable-timing in Singapore English. *Language and Speech* 43, 377-401.

[10] Maddieson, I. 2010. Correlating syllable complexity with other measures of phonological complexity. *Phonological Studies* 13, 105-116.

[11] Nazzi, T., Bertoncini, J., Mehler, J. 1998. Language discrimination by newborns: Toward an understanding of the role of rhythm. *J. of Exp. Psychology* 24, 756-766.

[12] Nolan, F., Asu, E.V. 2009. The pairwise variability index and coexisting rhythms in language. *Phonetica* 66, 64-77.

[13] Patel, A.D. 2008. *Music, Language, and the Brain*. Oxford and New York: Oxford University Press.

[14] Patel, A.D., Daniele, J.R. 2003. An empirical comparison of rhythm in language and music. *Cognition* 87, B35-B45.

[15] Patel, A.D., Iversen, J.R., Rosenberg, J.C. 2006. Comparing the rhythm and melody of speech and music: The case of British English and French. *J. Acoust. Soc. Am*. 119, 3034-3047.

[16] Ramus, F., Dupoux, E., Mehler, J. 2003. The psychological reality of rhythm classes: Perceptual studies. *Proc. 15th ICPhS* Barcelona, 337-342.

[17] Ramus, F., Mehler, J. 1999. Language identification with suprasegmental cues: A study based on speech resynthesis. *J. Acoust. Soc. Am.*105, 512-521.

[18] Ramus, F., Nespor, M., Mehler, J. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* 72, 1-28.