Prediction of Non-Linguistic Information of Spontaneous Speech from the Prosodic Annotation: Evaluation of the X-JToBI system

Kikuo Maekawa
Dept. Corpus Studies, National Institute for Japanese Language and Linguistics
10-2, Midori-cho, Tachikawa-shi, Tokyo 192-8561 JAPAN
E-mail: kikuo@ninjal.ac.jp

Abstract

Speakers' gender and age-group were predicted using the symbolic information of the X-JToBI prosodic labelling scheme as applied to the Core of the Corpus of Spontaneous Japanese (44 hours, 155 speakers, 201 talks). The correct prediction rate of speaker gender by means of logistic regression analysis was about 80%, and, the correct discrimination rate of speaker age-group (4 groups) by means of linear discriminant analysis was about 50 %. These results, in conjunction with the previously reported result of the prediction experiment of 4 speech registers from the X-JToBI information, shows convincingly the superiority of X-JToBI over the traditional J_ToBI. Clarification of the mechanism by which gender- and/or age-group information were reflected in the symbolic representations of prosody largely remains as open question, although some preliminary analyses were presented in the current paper.

Keywords: Corpus of Spontaneous Japanese, X-JToBI, non-linguistic information, age, gender

1. Aim of the study

It is widely recognized that speech signal conveys, in addition to linguistic information, various non-linguistic information about the speaker's physical status like gender and age. The recognition of non-linguistic information can often be important for the understanding of the message generated by the speaker, and/or, the proper management of discourse.

Needless to say, acoustic properties of speech signal like speech fundamental frequency (F0) and formant frequencies provide powerful cues for the recognition (see the discussion in section 4). But these are by any means the only cues in human recognition. There seems to be symbolic cues for the recognition like the choice of phrase final intonations (BPM, see below) or complex manipulations of prosodic boundaries (BI, see below).

Because these symbolic cues can't be extracted automatically from the speech signal at the present state of speech processing, it is important to examine if it is possible to predict the non-linguistic information from the symbolic prosodic annotation given to a speech corpus.

The primary aim of the present study consists in the evaluation of the predictability of speakers' gender (sex) and age from the relativized frequency data of the labels in a prosodic annotation scheme known as the X-JToBI scheme (Maekawa et al., 2002) as applied to the Core of the Corpus of Spontaneous Japanese (Maekawa et al., 2000, Maekawa 2003). The secondary aim of the study consists in the comparison between the traditional J_ToBI (Venditti, 1997) and the augmented X-JToBI schemes with respect to their ability to predict non-linguistic information.

2. Data

As shown in Table 1, the CSJ-Core consists of 201 speech files of about 44 hours long spoken by 155 different speakers, and covers 4 speech registers. APS (or academic presentation speech) is live recording of academic presentations covering the meetings of engineering, humanities, and social sciences. SPS (or simulated public speaking) is extemporaneous speech on everyday topics by recruited layman subjects. The topics of SPS include, for example, "the town where I live", "the most joyful/saddest memory of my life" and so forth. In addition to these monologue speeches, small amount of dialogue and reproduction speeches were included in the CSJ-Core for the sake of investigating phonetic and/or linguistic differences between monologue and dialogue on the one hand, and spontaneous and read speeches on the other. Most of the dialogues are interviews concerning the contents of an APS or SPS. Only the speech of interviewee, i.e., the original speaker of the APS or SPS, is analyzed. By reproduction speech is meant reading aloud of the transcription of an APS or SPS done by the same speakers.

REGISTER	N OF SPEECH*	TOTAL HOUR
APS	24 / 46	18.7
SPS	54 / 53	19.9
Dialogue	9 / 9	3.7
Reproduction	3 / 3	2.1

Table 1: Registers of the CSJ-Core
* Numbers to the left and right of a slash stand for female and male speakers, respectively.

REGISTERS	1930s*	1940s	1950s	1960s	1970s
APS	0/1†	2/3	3/7	11/13	8/22
SPS	5/5	5/4	11/10	16/17	17/17
Dialogue	0	0	0	3/6	6/3
Reproduction	0	0	0	1/1	2/2

Table 2: Distribution of the age of speakers
* "1930s" includes speakers born before1930. Also, "1970s" includes speakers born after 1980.
† Numbers to the left and right of a slash stand for female and male speakers, respectively.

Table 2 shows the distribution of speakers' birth year per decade. As can be seen from the table, speakers of dialogue and reproduction speech are concentrated in the 1960s and 1970s.

All speeches in the CSJ-Core were annotated in terms of segmental and prosodic characteristics using the X-JToBI annotation scheme, which is an extension for spontaneous speech of the original J_ToBI. Among the 6 tiers (word-, segment-, tone-, BI-, prominence-, and, miscellaneous-tiers) of the X-JToBI annotation, 4 tiers are of special interest for prosodic labeling.

Table 3 lists the main labels used in the 4 tiers and their frequencies in the CSJ-Core. Labels augmented in the X-JToBI extension are shown by an asterisk in the second column. Detailed explanations of the X-JToBI labels are omitted due to space limitation, but glosses are shown in the last column of the table. The frequency information was relativized by dividing the absolute frequency of a given label in a speech file by the total number of accentual phrases (AP) contained in the speech. In addition, the mean speaking rate (SR) was computed for each AP. The unit of SR is [mora/sec].

TIER	AUG.	LABEL	N	SYNOPSIS
Tone		L%	122675	Falling BPM
		H%	31115	Rising BPM
		HL%	10636	Rising-Falling BPM
	*	HLH%	14	Rising-Falling-Rising BPM
	*	LH%	419	"Insisting rise" BPM
	*	L%>	2143	Prolongation of L% tone
	*	H%>	3023	Prolongation of H% tone
BI	*	1+p	5864	Word boundary followed by a pause
		2	55252	Ordinary AP boundary
	*	2+p	9519	AP boundary followed by a pause
	*	2+b	9226	AP boundary followed by a BPM
	*	2+bp	4655	AP boundary followed by a pause and a BPM
		3	91373	IP boundary
	*	W	131	Words with multiple lexical accents
	*	P	1044	Word-internal pause
	*	PB	1186	Parasitic prosodic boundary
	*	F	36283	Filled pause
	*	D	6358	Fragmented word
Prominence	*	PNLP	1162	Penult Non-Lexical Prominence
	*	FR	3185	"Floating rise" variant of H%
	*	HR	215	"Hooked rise" variant of H%
	*	EUAP	2214	Emphasized Unaccented Accentual Phrase
Miscellaneous	*	QQ	250	Quasi-Question

Table 3: List of main X-JToBI labels

3. Analysis

3.1 Prediction of speakers' gender

Speakers' gender was predicted by means of logistic regression analysis using the glm( ) function of the stats library of the R language (version 2.14.1) using 4 different sets of independent variables.

Results were summarized in Table 4. The second row of the table shows the case when all independent variables were used for prediction. The third column shows the case when the variable of speaking rate was removed from the set of independent variables. In the fourth column, variables concerning disfluency (namely, F and D) were further removed from the set. The fifth column is concerned with the prediction using only the traditional J_ToBI variables. And in the last column were shown the results of Welch t-test as applied to each variables.

Each row of Table 4 shows the significance of each variables in the logistic regression analyses and t-tests. Blank row means that the variable was not significant at the level of 0.05. The rows marked with "---" were not involved in the regression analyses. The last row of the table shows the rates of correct prediction (in percentage) for each of the 4 prediction conditions as mentioned above.

Correct prediction rate was the highest when all variables, namely all X_JToBI labels and SR was used as independent variables, but the performance of the analysis using only X-JToBI variables was nearly as good as the analysis involving SR.

On the other hand, the performance of analysis using only J_ToBI variables (namely, "L%", "H%", "HL%", "2", and "3") was much worse than the previous ones. Actually, the performance of J_ToBI variables was not distinctly higher than the chance level, i.e., the case when all speakers were predicted as female (namely

111/201*100=55.2%).

Variables	XJToBI with SR	X_JToBI	X_JToBI wo F,D	J_ToBI	T test
L%					
H%		.			
HL%					
HLH%				--	
LH%				--	
L%>				--	
H%>				--	
1+p				--	
2	**	**	**		
2+p	*	*	.	--	.
2+b	**	**	*	--	
2+bp	**	***	***	--	.
3	**	**	*		
W				--	
P		**	**	--	.
PB			.	--	***
F	**	**	--	--	***
D			--	--	
PNLP		*	**	--	***
FR	.	**	**	--	***
HR	**	**	**	--	***
EUAP	.			--	
QQ				--	***
SpkRate	***	--	--	--	***
Correct Prediction Rate (%)	83.1	80.1	79.1	58.7	---

*** 0, ** 0.001, * 0.01, . 0.05, -- Not Available

Table 4: Results of logistic regression analyses using 4 different sets of independent variables and t-test.

Variables	XJToBI with SR	X_JToBI	X_JToBI wo F,D	J_ToBI	ANOVA
L%	-1.31	-1.49	-1.49	0.39	
H%	-1.88	-2.05	-1.91	-1.14	
HL%	-1.81	-1.94	-1.95	-0.61	
HLH%	0.98	1.00	0.95	--	
LH%	-0.91	-1.17	-1.46	--	*
L%>	-0.64	-0.57	-0.09	--	.
H%>	-1.38	-1.49	-1.61	--	
1+p	-1.62	-1.42	-0.81	--	.
2	-0.05	-0.19	0.14	-2.22	
2+p	-0.35	-0.44	-0.21	--	
2+b	0.18	-0.01	0.14	--	.
2+bp	3.08	3.33	3.25	--	
3	0.03	0.07	0.51	-0.57	
W	1.36	1.51	1.23	--	
P	-1.37	-1.06	-0.17	--	
PB	1.36	1.61	0.95	--	
F	1.70	1.47	--	--	
D	2.39	2.54	--	--	
PNLP	1.91	1.86	2.98	--	.
FR	-3.82	-4.33	-4.03	--	***
HR	-2.25	-2.08	-2.31	--	***
EUAP	0.82	0.82	0.39	--	
QQ	0.96	0.95	1.37	--	
SpkRate	-1.58	--	--	--	*
Correct Prediction Rate (%)	59.3	55.9	55.4	32.2	---

*** 0, ** 0.001, * 0.01, . 0.05, -- Not Available

Table 5: Results of ordered logistic regression analyses and one-way ANOVA (See text).

3.2 Gender-sensitive variables

Although it is not the aim of the present study to analyse extensively the way each X_JToBI variable contributes to the predictions, interesting cases are shown in figure 1 that shows box-whisker plots of some gender-sensitive X_JToBI variables where Welch t-test showed $p<.0001$ significance in the comparison between the male and female speakers (see the last column of Table 5).

There is trading-relation among the values of "2", the sum of the values of "2+p", "2+b", and "2+bp", and the value of "3" as shown in figure 2. This is because all these values are concerned with the classification of the strength of AP (accentual phrase) boundary. The sum of the raw occurrence frequencies of these labels is identical to the number of authentic APs in the whole data. As can be seen from figure 2, male speakers tends to use more "2" boundary rather than other boundaries of "2+" class.

The variable "PB," or "Parasitic Boundary", is a special break index (BI) applied to the cases where the end of an AP is associated with more than 2 final boundary tones, the typical case being L%H% boundary followed by another H% tone.

Many "PB" boundaries occur when speakers use so-called "Quasi-Question" as represented by the "QQ" label, which stands for the cases where an utterance that is interpreted pragmatically as an ordinary statement while the end of utterance is associated with a yes-no-question like rising intonation. It is important to note here that "QQ" can be regarded to be the Japanese counterpart of English "high rising terminal" or "uptalk", and, as in English, it is used mostly by female speakers of various age-groups. In fact, as shown in figure 1, "QQ" occurs almost exclusively in the speech of female speakers in the current data.

"HR", a special variant of rising intonation known as the "hooked rise" (Kawakami, 1963), also showed a distribution strongly skewed toward the female speakers.

The relativized occurrence rates of "F", "PNLP", and "FR", on the other hand, are higher in males' speech rather than in females'. And, males have higher speaking rate than females.

3.3 Prediction of speakers' age-group

The second analysis is concerned with the prediction of speakers' age-groups (Table 2). Because the distribution of the speakers' age in the CSJ-Core is strongly skewed in dialogue and reproduction registers, speech data belonging to these registers were removed from the analysis. The resulting data consisted of 177 speeches of APS and SPS.

Table 2 also shows that relatively fewer number of subjects belong to the groups of 1930s and 1940s. To correct this, subjects belonging to 1930s and 1940s were merged into a single age-group. As the result, speakers were classified into 4 age-groups; "=<1940s", "1950s", "1960s", and "1970s=<", from the most elderly to the

youngest.

The age-groups of the speakers of 177 monologues were predicted by means of ordered logistic regression analysis (proportional odds logistic regression) using the polr( ) function of the MASS library of the R. The results are summarized in Table 5, where the values shown in the second to fifth columns are the t-values (namely, the estimated regression coefficient divided by the standard error). The last column summarizes the result of one-way ANOVA using the oneway.test( ) function of the R.

And, in the last row of the table are shown the rates of correct prediction. Prediction using the whole X-JToBI variables and SR (speaking rate) achieved the highest prediction rate, but the performances of the X-JToBI variables per se and that of X-JToBI variables without F and D were not much behind.

The performance of the traditional J_ToBI variables was, on the other hand, distinctively behind the predictions using the X-JToBI variables. As a matter of fact, its mean correct prediction rate of 32.2% is lower than the chance level (the case when all subjects are classified as belonging to the age-group of "1970s" or younger", i.e., 64/177*100= 36.15%).

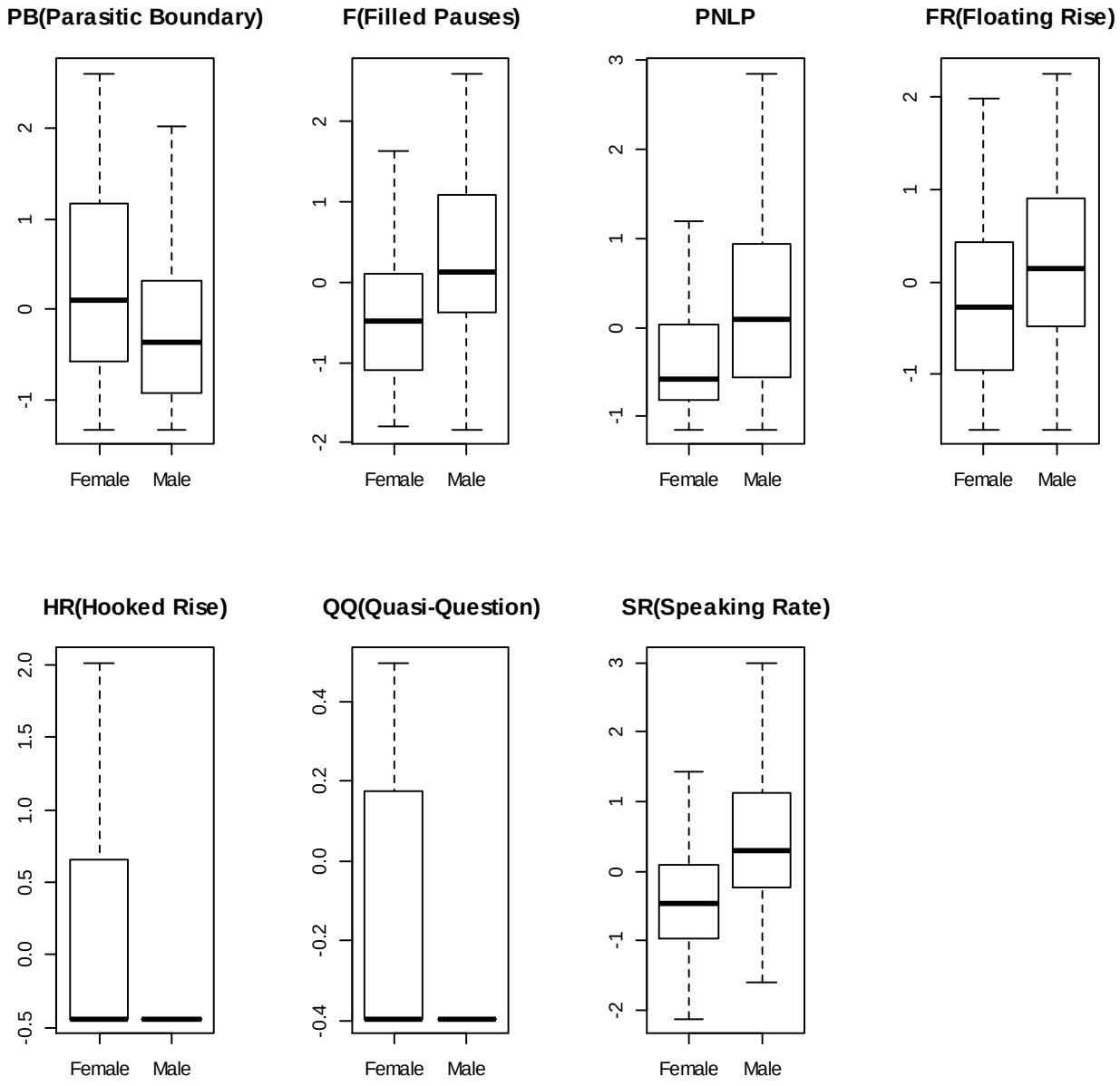

Figure 1: Box-whisker plots of some gender-sensitive X-JToBI variables.
The ordinates are standardized z-scores (see text).

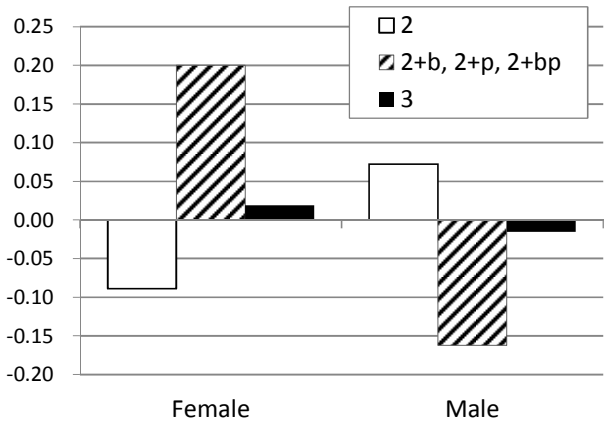

Figure 2: Comparison of the mean relativized frequencies of AP boundaries between male and

3.4 Age-sensitive variables

Figure 3 is the box-whisker plots of some age-sensitive X-JToBI variables as a function of speakers' age-group. Variables that showed higher-than-0.05 significance in the ANOVA part of Table5 were selected. Linear trend was observed in variables "1+p", "2+b", "PNLP", "FR", "HR", and "SR."

The interpretation of the correlation shown in figure 3 is not easy, but here are some pilot interpretations. It is well known that, generally, younger speakers speak faster than elderly speakers (see the panel of "SR").

The panel of "2+b" suggests that younger people use more BPM (Boundary Pitch Movement, a local intonation marking the end of an AP). "PNLP", "FR", and "HR" are all concerned with BPM, but "PNLP" and other 2 BPMs showed opposite correlation patterns. Younger speakers used less "PNLP" than elderly speakers. It may be because younger speakers were not good at producing logically constructed monologues, because the primary linguistic function of a PNLP was to prvide a cue for topic segmentation (Maekawa, 2011a).

On the contrary, younger speakers use the BPM like "FR" and "HR" more frequently than elderly speakers. Probably, this was because younger speakers tended to produce their monologues more-or-less 'emotionally' rather than 'logically'. Use of BPMs like "FR" and "HR" are known to provide cues for the speakers' attitudes or intentions (Kawakami, 1963).

Interpretation of the variable "LH%", "L%>", and "1+p" are difficult. It should remain as an open question at the present stage of inquiry.

4. Discussion and conclusion

It is widely acknowledged that non-linguistic information in speech was transmitted by acoustic cues like speech fundamental frequency (f0), formant frequencies, and voice-source characteristics.

The results of the current study revealed, however, non-linguistic information like gender and age could be transmitted by prosodic characteristics that are symbolic, as well. Moreover, the performance of the prediction by means of symbolic variables (about 80% in the case of speakers' gender) is nearly in the same level as the one reported in Schuller et al (2010) who used continuous acoustic parameter for prediction (but we have to be careful about the direct comparison, because the task used in the latter study was quite different from the one reported in the current paper).

It is not clear, at the present stage of the study, if native speakers of Japanese are deliberately using symbolic cues for the transmission of gender- and/or age-information. But it seems to be probable that speakers are using symbolic cues for the perception of non-linguistic information in spontaneous Japanese.

The second contribution of the current study consists in the confirmation of the superiority of the X-JToBI system over the traditional J_ToBI for the prediction of non- and paralinguistic information in speech.

Our previous study showed that it was possible to automatically discriminate the 4 speech registers (APS, SPS, dialogue, and reproduction speech) of the speech files of the CSJ-Core with higher than 85% accuracy (in the case of closed-data, Maekawa, 2011b). The results reported in the present study strongly reinforce the conclusion of the previous study.

5. Acknowledgements

Parts of this study are supported by the grant-in-aid for scientific research from the JSPS to the present author (No 23520483) and the research budge of the National Institute for Japanese Language and Linguistics (Project title "Foundation of corpus Japanese linguistics").

6. References

Kawakami, S. (1963). "Bunmatsu nadono jooshoochoo nitsuite." *Kokugo Kenkyuu*, 6, 21-31.

Maekawa, K (2003). "Corpus of Spontaneous Japanese: Its design and evaluation." *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, 7-12.

Maekawa, K. (2011a). "Phonetic Shape and Linguistic Function of Penultimate Non-Lexical Prominence." *Journal of the Phonetic Society of Japan*, 15 (1), 16-28,

Maekawa, K. (2011b) "Discrimination of speech registers by prosody." *Proc. 17th ICPhS*, 1302-1305.

Maekawa, K., H. Kikuchi, Y. Igarashi and J. Venditti (2002). "X-JToBI: An extended J_ToBI for

spontaneous speech." *Proc. ICSLP2002*, 1545-1548.
Maekawa, K., H. Koiso, S. Furui and H. Isahara (2000). "Spontaneous speech corpus of Japanese." *Proc. LREC 2000*, 947-952.
Schuller, B. et al. (2010). "The INTERSPEECH 2010 Paralinguistic Challenge." *Proc. INTERSPEECH 2010*, 2794-2797.
Venditti, J. (1997). "Japanese ToBI Labelling Guidelines." *Ohio State University Working Papers in Linguistics*, 50, 127–162.

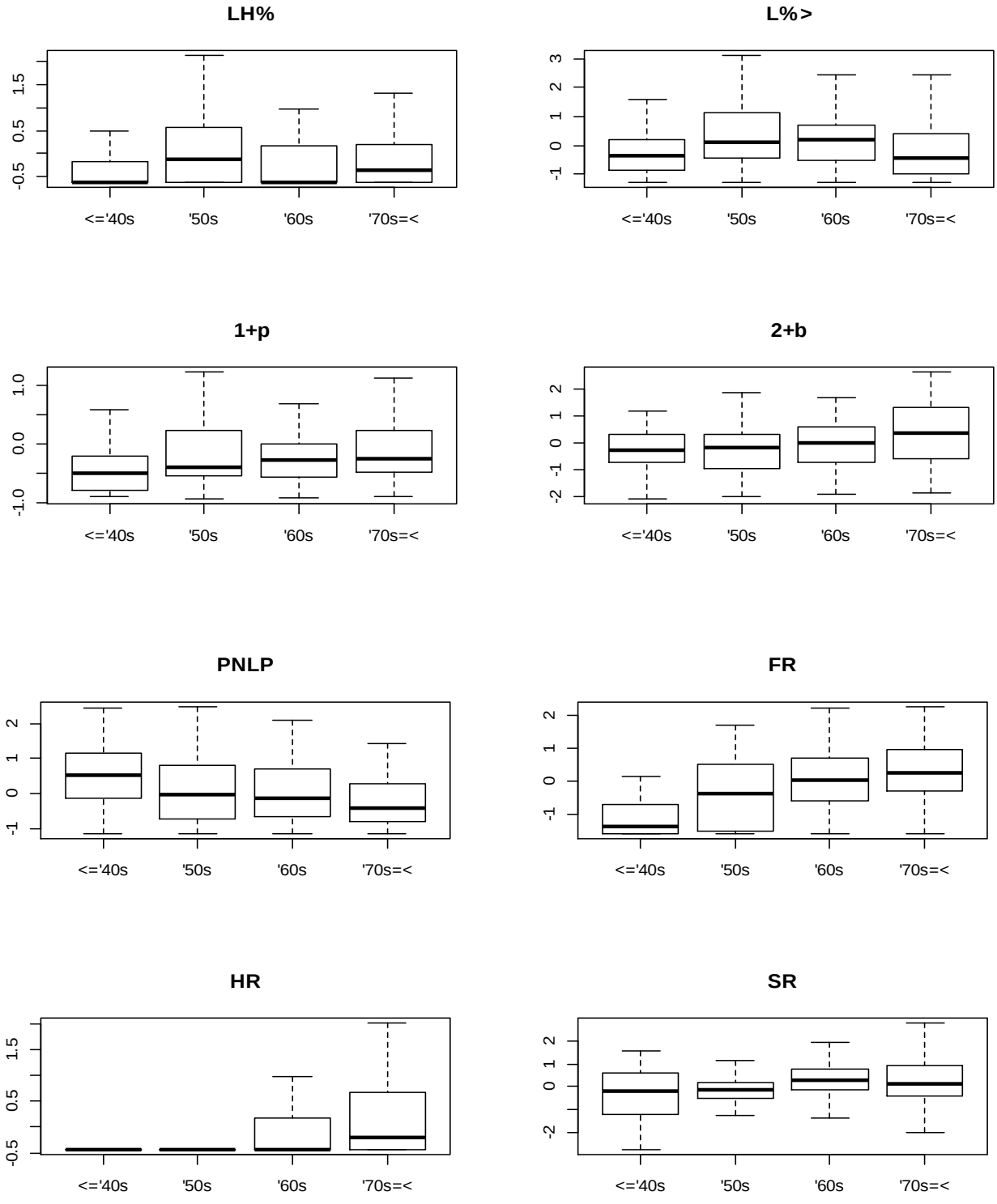

Figure 3: Box-whisker plots of some gender-sensitive X-JToBI variables.
The ordinates are standardized z-scores (see text).