DISCRIMINATION OF SPEECH REGISTERS BY PROSODY

Kikuo Maekawa

Department of Corpus Studies, National Institute for Japanese Language and Linguistics, Japan
kikuo@ninjal.ac.jp

ABSTRACT

Normalized frequency data of X-JToBI prosodic labels were used to automatically discriminate 4 speech registers –academic presentation, simulated public speaking, dialogue, and reproduction speech– of the *Corpus of Spontaneous Japanese* (CSJ). It turned out that the use of prosodic label frequency information and speaking rate could achieve more than 85% accuracy (closed data). It also turned out that the prosodic cues contributing to the classification were distributed pervasively throughout speech.

Keywords: CSJ, spontaneous speech, X-JToBI, register, prosody

1. INTRODUCTION

It is widely acknowledged that people change the phonetic shape of their utterances depending on the social settings in which their speeches are generated. This kind of variation has been studied extensively in the variationist sociolinguistics [3].

From a phonetician's point of view, however, existing studies of this field share one important problem: they are concerned almost exclusively with the contrastive and/or segmental aspects of speech and do not pay enough attention for the phrasal and/or prosodic aspects.

It is widely believed not only by laymen but also by specialists that, in the actual life, native speakers of a language change the prosody of their speech depending on the register of speech, and are very keen in perceiving such characteristics in the speech of someone else [1].

This belief, however, has not been studied in a scientific way, with the exception of Nevalainen's analysis of the prosodic annotation of the London-Lund corpus [7] that will be discussed later.

The aim of the present study is to show the presence of systematic correspondence between the registers of spoken texts recorded in a corpus of spontaneous Japanese speech and the frequency distribution of the labels used in the prosodic annotation of the corpus.

2. DATA

2.1. The corpus

The data analyzed in this study consists of the phonetically annotated part of the *Corpus of Spontaneous Japanese* [4, 5], which is known as the CSJ-Core. 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.

Table 1: Properties of the CSJ-Core.

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

* Numbers to the left and right of a slash stand for male and female speakers, respectively.

As can be read from the table, 87% of the CSJ-Core is occupied by APS and SPS. This was because CSJ was designed primarily as a resource for machine learning of acoustic- and language-models for a new-generation automatic speech recognition system that can handle more-or-less spontaneous monologue [2].

A 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 APS or SPS. Only the speeches of interviewees, i.e., the original speakers of APS or SPS, are analyzed. By reproduction

speech is meant reading aloud of the transcription of an APS or SPS made by the same speakers.

Table 2: Main X-JToBI labels.

TIER	AUG.	LABEL	N
Tone		L%	97,556
		H%	24,621
		HL%	8,863
	*	HLH%	8
	*	LH%	308
	*	L%>	480
	*	H%>	2,275
BI	*	1+p	3,872
		2	42,568
	*	2+p	7,155
	*	2+b	7,098
	*	2+bp	3,456
		3	71,383
	*	W	35
	*	P	263
	*	PB	1,033
Prominence	*	PNLP	856
	*	FR	2,535
	*	HR	207
	*	EUAP	1,667
Miscellaneous	*	QQ	220

2.2. Variables for statistical analyses

All speeches in the CSJ-Core were annotated in terms of segmental and prosodic characteristics using the X-JToBI annotation scheme [6], which is an extension for spontaneous speech of the original J_ToBI [9]. 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 2 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. Definitions of each of the X-JToBI labels are omitted in the present paper due to the space limitation. But some of them are explained briefly in section 3.2 below. See literature [6] for the definitions of X-JToBI labels.

In addition to the X-JToBI labels, information about the mean speaking rate was used in the statistical analyses. Mean speaking rate (SR) was computed for each accentual phrase (AP), the unit for SR being [mora/sec].

All frequency information of the X-JToBI labels was normalized by dividing them by the number of APs comprising the speech file, while SR was not normalized. Lastly, all variables including SR are z-transformed.

3. ANALYSIS

3.1. LDA using the whole data

The normalized data was analyzed by means of linear discriminant analysis (LDA). To begin with, an LDA was conducted using all X-JToBI labels as independent variables and nothing else. The lda() function in the MASS library of the R language (ver. 2.10.1) was used. Table 3 shows the prediction results. Rows and columns of the table correspond respectively to the correct and predicted registers. Symbols of "A", "D", "R", and "S" are used to refer to the registers of APS, dialogue, reproduction, and SPS. The total correct classification rate was 85.1%, and the result of leave-one-out cross validation was 78.1%. These two types of correct classification rates will be referred to, respectively, as the 'closed-data' and 'CV' hereafter. Figure 1 shows the distribution of all CSJ-Core speeches on a plane generated by the first 2 discriminant functions (LD1 and LD2) using all X-JToBI variables and SR.

Table 3: Prediction result of LDA.

	A	D	R	S
A	56	0	0	14
D	0	12	0	6
R	2	0	3	1
S	8	0	1	98

Figure 1: Distribution of the whole CSJ-Core speeches on the LD1-LD2 plane. All X-JToBI labels and SR were used.

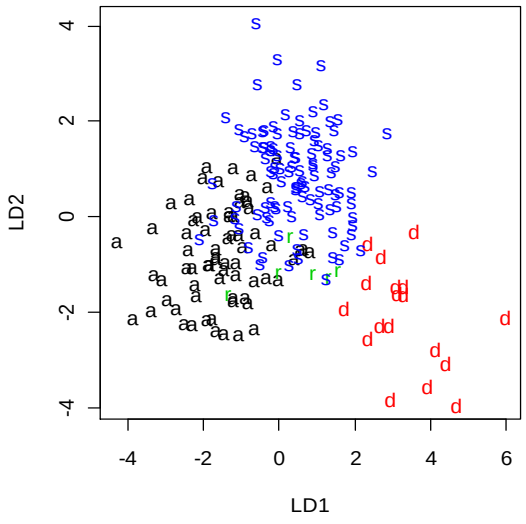


Table 4: Comparison of the contributions of X-JToBI tiers.

VARIABLES	CLOSED-DATA[%]	CV[%]
All variables (Tbl.3)	85.1	78.1
Tone	70.1	68.2
BI	75.6	74.6
Prominence	71.6	68.2
Tone+BI	80.6	78.1
Tone+Prominence	76.1	72.6
BI+Prominence	79.1	78.6

Table 4 compares a series of LDAs using various subsets of the X-JToBI labels as the independent variables. The first row shows the results of Table 3, the second, third, and fourth rows show the results of 3 separate LDAs using the information involved in each single X-JToBI tier. The remaining 3 rows show the results of the combinations of two tiers.

As for the contribution of a single tier, it was the variables involved in the BI tier that achieved the best correct classification rates in terms of both closed-data and CV, but the achievements of the Tone and Prominence tiers were no less important. As for the combination of two tiers, pairs of Tone and BI, and that of BI and Prominence showed the best performance. In these pairs, the results of CV were as high as or higher than that of the complete X-JToBI data.

3.2. Tests of individual variables

The contribution of each independent variable was examined individually. For this purpose, first, a series of Kruskal-Wallis tests were applied for the variables shown in Table 1 and SR. Variables that showed p<.01 significance included 1+p (word boundary followed by a pause), 2+b (AP boundary followed by a boundary pitch movement, or BPM), 2+bp (AP boundary followed by a BPM and a pause), 3, PB (parasitic prosodic boundary, i.e., two consecutive boundary tones), L%, H%, HL%, L%> (prolonged L% boundary tone), H%> (prolonged H% tone), EUAP (emphasized unaccented AP with strong pitch range reduction in the following AP), FR (‘floating rise’ variant of the L%H% or L%HL% BPMs), PNLP (penultimate non-lexical prominence, a variant of L%HL% BPM), QQ (‘quasi-question’, Japanese counterpart of English ‘uptalk’), and SR.

Subsequently, post-hoc tests were applied for these variables for all pairs of registers. Table 5 summarizes the results. Column names like “A/D” and “R/S” stand respectively for the pairs of “APS and dialogue” and “Reproduction and SPS.” Symbols showing the significance levels are: *** for p<.001, ** for p<.01, * for p<.05, and, - for p>=.05. Table 5 shows that at least 2 variables are statistically significant for any pair of registers.

Table 5: Summary of the post-hoc tests.

VAR	A/D‡	A/R	A/S	D/R	D/S	R/S
1+p	-	-	**	-	-	-
2+b	***	-	***	-	-	-
2+bp	**	-	***	-	-	-
3	***	-	***	-	-	-
PB	*	-	***	-	***	**
L%	-	-	***	-	-	-
H%	-	-	***	-	-	-
HL%	-	-	***	*	-	***
L%>	-	-	**	**	-	-
H%>	-	**	--	*	-	*
EUAP	**	-	--	-	-	-
FR	-	-	**	-	-	-
PNLP	***	-	**	-	-	-
QQ	-	-	**	-	-	-
SR	***	*	***	-	***	-

3.3. LDA with variable selection

Lastly, to evaluate the relative importance of X-JToBI labels for the register discrimination, an LDA with stepwise variable selection was conducted using the sdis() function of R language written by Shigenobu Aoki [8]. The following variables were selected as the result of the forward and backward selection: HL%, L%>, 1+p, 2+p, 2+b, PB, P and SR. It is interesting that none of these variables belongs to the original J-ToBI labels with the sole exception of HL% (See the column AUG of Table 2).

To check the validity of this selection, ordinary LDA was conducted using only the variables selected in the previous analysis. The resulting correct classification rates were 85.1% in the closed-data, and 81.1% in CV. This performance was at virtually the same level as the best performance shown in Figure 1, where all X-JToBI labels and SR were used as variables.

Variable selection was conducted using only the X-JToBI variables also. This time, the following variables were selected: HL%, L%>, 1+p, 2+p, 2+b, 3, PB, P, and PNLP. The performance of the LDA using only these variables was 80.6% and 80.1%, respectively, for closed-data and CV.

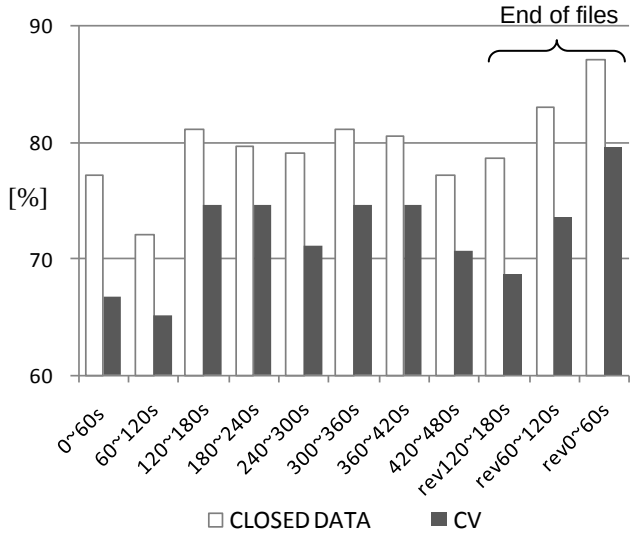
3.4. LDA using partial data

In the analyses reported in the previous sections, a speech was characterized by the frequencies of all labels that occurred in a speech file. How much does correct classification rate degenerate if the data is sampled from a limited part of the file? To answer this question, Figure 2 compares the results of a series of LDAs using the X-JToBI labels (no SR) involved in a rectangular time window of 60 seconds long that shifts its starting location from 0 to 480 seconds by the interval of 60 seconds.

The last three pairs of bars in the figure, on the other hand, show the results of LDAs when the same time window is located near the end of a speech file. The symbol "rev60~120", for example, represents the case where the beginning and end of a window was located 120 seconds and 60 seconds respectively from the end of a file.

Note also that the data of low frequency labels (LHL% and H%>) were excluded from the analysis, as there were cases where the labels did not occur within the time-window for analysis.

Figure 2: Results of LDA with a data-window of 60 sec.



Correct classification rates of closed-data and CV were both low at the beginning, and high at the end, but in the intermediate window locations, it is difficult to find any trend between the location of the data-window and the correct classification rates. Roughly speaking, the achievements of closed-data and CV are scattered around the values of 80% and 70% respectively.

4. CONCLUSION

Results reported in this paper lend strong support for the widespread but unverified belief that prosody can provide strong cues for the human discrimination of registers. As for English, Nevalainen reported statistical analysis of the prosodic labels of the London-Lund corpus in relation to discourse types [1], but her analysis was concerned only with nuclear tones, and, the automatic discrimination of speech samples was not attempted.

It is the new finding of the present study that more than 80% accuracy could be achieved in the automatic classification of speech registers by use of simple frequency information of prosodic labels.

Another new finding is the pervasiveness of the distribution of prosodic features characterizing speech registers. The finding that speech specimen of 60 second long is enough for the 80% correct characterization of speech register seems to be in congruence with our intuition about the role of prosody in spontaneous speech communication.

Lastly, it turned out that the labels augmented in the X-JToBI system played important role for the discrimination. This fact suggests strongly the effectiveness of the X-JToBI annotation for the analysis of spontaneous Japanese.

5. REFERENCES

[1] Crystal, D., Davy, D. 1969. *Investigating English style.* London: Longman.

[2] Furui S., Maekawa, K., Isahara, H. 2000. Toward the realization of spontaneous speech recognition: Introduction of a Japanese priority program and preliminary results. *Proc. ICSLP2000*, 518-521.

[3] Labov, W. 1972. *Sociolinguistic Patterns.* University of Pennsylvania Press.

[4] Maekawa, K. 2003. Corpus of spontaneous Japanese: Its design and evaluation. *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, 7-12.

[5] Maekawa, K. 2010a. Coarticulatory reinterpretation of allophonic variation: Corpus-based analysis of /z/ in spontaneous Japanese. *Journal of Phonetics* 38(3), 160-174.

[6] Maekawa, K., Kikuchi, H., Igarashi, Y., Venditti, J. 2002. X-JToBI: An extended J_ToBI for spontaneous speech. *Proc. ICSLP2002*, 1545-1548.

[7] Nevalainen, T. 1992. Intonation and discourse type. *Text* 12(3), 397-427.

[8] sdis function of R language written by Shigenobu Aoki *http://aoki2.si.gunma-u.ac.jp/R/sdis.html*

[9] Venditti, J. 1997. Japanese ToBI labelling guidelines. *Ohio State University Working Papers in Linguistics* 50, 127-162.