Performance of Segmental and Prosodic Labeling of Spontaneous Speech

KIKUCHI Hideaki†‡ MAEKAWA Kikuo†
†National Institute for Japanese Language
‡Waseda University
{kikuchi,kikuo}@kokken.go.jp

ABSTRACT

In an attempt to construct a large-scale database of spontaneous speech, the authors planned to give segmental and prosodic labels to spontaneous Japanese speech. In this paper, the performance of those labeling will be reported. First, the performance of automatic segmental labeling by Hidden Markov Model was verified. Sample speech of about four hour long was automatically phoneme labeled and compared to the results of hand-labeling. It turned out that average of label boundary difference with hand labeled data was 14.3[ms]. Second, the performance of prosodic labeling by newly proposed labeling scheme named X-JToBI(eXtended J_ToBI) was verified. The analysis of labeled data showed that newly added inventories appeared in the data of spontaneous speech and rate of inter-labeler agreement increased in nearly all types of labels.

1 INTRODUCTION

Study of spontaneous speech requires large database because spontaneous speech is inherently more variable than read or laboratory speech. Such a database must be annotated with segmental and prosodic labels. Since 1999, the authors have been involved in the compilation of a large-scale corpus of spontaneous speech known as the *CSJ*, or Corpus of Spontaneous Japanese [1]. This corpus involves the digitized speech, transcribed speech, POS annotation of about 650 hour spontaneous speech corresponding to about 7 million morphemes. In addition, we will provide segmental labels and prosodic labels for a true subset of the *CSJ*, called the Core, containing about 45 hour speech, or 500 thousand morphemes.

It is generaly believed that labeling spontaneous speech is more difficult than labeling read speech because of wide variety of acoustic and linguistic features. The aim of this paper consists in making report of problems in labeling spontaneous speech and verification of effectiveness of our answers to those probelems.

2 SEGMENTAL LABELING

Segmental, or phoneme, label is indispensable when we conduct a corpus-based phonetic analyses and/or speech modeling. However, the cost of manual labeling becomes too high and unaffordable as the size of the corpus becomes bigger. Hence arose the necessity for the accurate and efficient method of automatic labeling. So far, methods based upon viterbi alignment using HMM phoneme model were proposed and turned out to be rather effective [2]. In our work, the first step consisted in the automatic conversion of the transcribed speech into phoneme labels. The phoneme labels were aligned to speech signal by an automatic phoneme labeling technique based upon the HMM phoneme model. Then, the phoneme labels were converted to segmental labels[1] and adjusted by hand. The inventory of segmental labels is not purely phonemic; some labels are phonetic, or sub-phonemic, representing events like phonetic palatalization, burst of stop consonants, devoiced vowels, and so on.

In the following sections, inter-labeler agreement of segmental labeling of the *CSJ* will be shown at first and then, performance of automatic labeling will be reported.

2.1 Inter-labeler reliability

A pilot experiment was conducted to know the inter-labeler difference of accuracy. Two labelers labeled *CSJ* speech samples of about 3 minutes long. In the hand labeling, merger of more than two labels into a single label was permitted. Merged labels were applied mostly for diphthongs and the fricative-devoiced vowel sequences. Table 1 shows the result. 'B.D.' denotes mean difference of boundary location between the labelers. The total numbers of labels, both simplex and merged, do not agree among labelers. The mean difference between the corresponding segment boundaries was about 8[ms]. This value can be considered to be the target value of automatic labeling. Because more than 15 hours of the Core has already been labeled manually, we use them as the 'correct' labels in the rest of this paper.

Table 1: Comparison between the results of segmental labeling by hand.

Speech		Labeler A		Labeler B		
ID	Time [sec]	Total	Merged	Total	Merged	B.D. [ms]
A	22.4	298	29	282	36	7.78
B	24.2	348	30	299	36	7.67
C	22.9	312	31	291	41	7.70
D	33.6	408	25	378	45	8.34
E	22.2	275	19	261	25	6.76
F	32.0	396	33	350	56	11.46
G	19.2	253	17	237	26	8.37
H	28.3	307	24	306	26	9.73
I	23.0	301	22	275	35	7.03
Whole		2988	230	2679	326	8.37
(Ratio[%])			(7.9)		(12.1)	

Table 2: Condition of acoustic analysis of models.

Sampling Frequency	16[kHz]
Window size	25[ms]
Window shift	10[ms]
Feature parameter	MFCC(12 dim.)+ΔMFCC(12 dim.) +ΔPower

2.2 Performance of automatic labeling

An automatic phoneme labeling experiment was conducted to know the basic performance of automatic labeling of spontaneous speech by HMM.

2.2.1 Phoneme Model

Two types of HMM were used as phoneme model: read and spontaneous speech models. The model delivered with the Japanese Dictation Kit[3] was used as the read speech model('IPA'). The amount of the training data for this model was about 40 hours. The spontaneous speech model('CSJ') was trained with the *CSJ* speech of about 59 hour[4]. Table 2 shows the condition of acoustic analysis used for these models.

2.2.2 Method

The evaluation data for the experiment was the 20 speech files (about 4 hours long) excerpted from the *CSJ*. This data was automatically labeled by HTK[5] in combination with various models trained with different training data and model parameters. Note, in passing, that evaluation data was included in the training data of spontaneous speech model. Difference of label boundaries were calculated for about 100,000 label pairs. Labels that do not correspond and merged labels were excluded from the calculation. In the rest of this paper, standard deviation of boundary difference is used as the criterion of labeling accuracy.

2.2.3 Result

Table 3 compares the performance of automatic labeling applied for read and spontaneous speech samples. ATR database of phoneme-balanced sentences[6] was used as the samples of read speech('ATR-DB'). 10 files of male speakers excerpted from the above-mentioned 20 *CSJ* files were used as the samples of spontaneous speech. 'APS' and 'SPS' are the two main speech types recorded in *CSJ*, namely, academic presentation speech (APS) and layman's simulated public speech (SPS) [1]. 'Mono' and 'tri' stands respectively for monophone and triphone model. Numbers in parenthesis show the numbers of states involved in each model. Number of Gaussian mixture is fixed to 16 for all models.

According to this table, performance of *CSJ* samples was worse than that of read speech. Judging from the data in the literature[7], which reported the results of automatic labeling of TIMIT dialogue speech by monophone model (22.7[ms] standard deviation), the result obtained with *CSJ* was intermediate between the read and dialogue speech.

Also, the monophone model trained with read speech('IPA-mono') showed the best performance (in terms of the standard deviation). Moreover, it was noteworthy that performance of triphone models was consistently lower than that of monophone models. Acoustic probability of triphone models is nonetheless higher than that of monophone models. This phenomenon has already been reported[7]. Better performance was obtained with context-independent model, because so-called 'glide' between the two consecutive phonemes was shared by the two adjacent phoneme models under the iterative training using fixed context (*i.e.* triphone context). The drawback of sharing glide features becomes apparent when the phoneme in question is marked by a sharp, abrupt acoustic changes as in the case of stop consonants.

It is also interesting to see that the performance of read speech models was better than that of spontaneous speech in terms of boundary position, while spontaneous speech models showed higher acoustic probability. The reason for this lies in the fact that the acoustic boundaries in the read speech data are clearer than that of spontaneous speech data, and, accordingly, the read speech models acquire the boundary features more accurately than the spontaneous speech models. We confirmed same tendencies in the result of labeling female speech by female phoneme model.

Table 3: Comparison of mean boundary difference[ms] by types of target data. (numbers in () stands standard deviation.)

Model	ATR-DB	APS	SPS
IPA-mono	-6.92(18.63)	-3.55(20.50)	-4.19(21.60)
IPA-tri(1000)	-7.18(20.22)	-4.55(21.53)	-4.49(22.38)
IPA-tri(2000)	-7.02(20.78)	-4.81(21.07)	-3.96(21.91)
IPA-tri(3000)	-8.32(20.78)	-5.81(21.64)	-4.87(22.00)
CSJ-mono	-9.33(20.43)	-6.44(21.30)	-7.17(22.23)
CSJ-tri(1500)	-6.99(19.94)	-4.05(21.43)	-4.45(22.91)
CSJ-tri(2000)	-7.19(19.92)	-3.80(21.33)	-4.19(22.76)
CSJ-tri(3000)	-7.11(20.07)	-3.91(21.59)	-4.20(22.84)

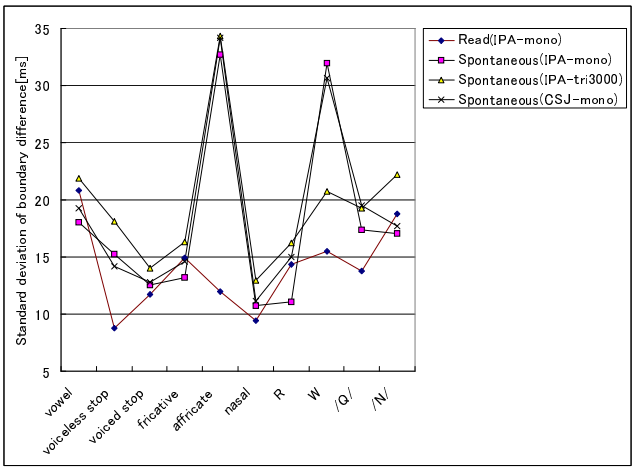

Fig. 1: Comparison of standard deviation of boundary difference by phoneme category.

Figure 1 compares the performance of read and spontaneous speech across acoustic models and phoneme categories. The performance of labeling spontaneous speech was worse in almost all phoneme categories. Especially, it was difficult to label affricate and glide(w) of spontaneous speech. Observing the results of hand-labeling of *CSJ*, affricate labels were not always divided into closure part and burst part. The ratio of merging these labels to the total number of affricate was 17.1% in preliminary labeling while the ratio of merger in all samples was about 10%. Affricate seems to have more variants than other categories and the variation becomes all the more complex in spontaneous speech. More elaborated handling of segmental unit is needed in order to improve the accuracy of spontaneous speech labeling.

3 PROSODIC LABELING

According to our pilot experiment inter-labeler reliability of J_ToBI[8] lowered considerably when the scheme was applied to spontaneous speech. To solve this problem, we proposed a new prosodic labeling scheme named X-JToBI[9], the extended version of J_ToBI[8]. Among the new characteristics of X-JToBI are 1) Exact match between the time-stamp of tone labels and the timing of physical events, 2) Enlargement of the inventory of boundary pitch movements, 3) Extension and ramification of the usage of break indices, and 4) Newly defined labels for filled-pause and non-lexical prominence.

In this section, frequency of newly introduced labels and the performance of the X-JToBI labeling is described.

3.1 Frequency of labels

Table 4 shows the frequency of tone labels of boundary pitch movement(BPM) in two types of monologue. The total amount of speech (excluding pauses) in the APS and SPS styles were 1.02 and 2.75 hours respectively. An expert labeler labeled all samples. It is intersting to see that speakers use L%H% more frequently in APS than in SPS, presumably to signal the so-called 'continuation rise'. It is also noteworthy that L%LH%, a newly introduced BPM, appeared across speech types.

Table 4: Frequency of tone labels of BPM
(numbers in () stands ratio in all BPMs.[%])

tone label	SPS		APS	
L%+H%	1683	(58.05)	1331	(78.80)
L%+HL%	1121	(38.67)	346	(20.49)
L%+LH%	95	(3.28)	12	(0.71)

Table 5: Frequency of BI labels
(numbers in () stands ratio in all BI labels.[%])

BI label	SPS		APS	
1	31696	(69.53)	8121	(55.73)
1+	42	(0.12)	9	(0.06)
1+p	254	(0.70)	164	(1.13)
2	4071	(7.82)	1333	(9.15)
2+	15	(0.03)	9	(0.06)
2+p	214	(0.41)	67	(0.46)
2+b	466	(0.90)	526	(3.61)
2+pb	48	(0.09)	8	(0.05)
3	7478	(14.37)	2617	(17.96)
D	280	(0.54)	97	(0.67)
D+	31	(0.07)	26	(0.18)
P	19	(0.04)	6	(0.04)
P+	12	(0.02)	5	(0.03)
<F	431	(0.83)	464	(3.18)
F	2426	(4.66)	1111	(7.62)
PB	61	(0.06)	10	(0.07)

Table 5 compares the frequency of the X-JToBI break index(BI) labels. The most salient difference between the two speech types was the higher relative frequency of '2+b' in APS. This label marks cases in which downstep is continued across one or more prosodic boundary marked with BPM(s). This particular intonation is found in the presentation of debutant researchers who seem to memorize his/her entire talk.

Newly introduced BI labels for disfluency(D,D+,P and P+) and filled-pause(<F and F) appeared frequently. From these results, we can say that newly introduced labels of tones and BIs are useful in the prosodic labeling of *CSJ*.

3.2 Accuracy and Inter-labeler reliablity

In this part, the accuracy and inter-labeler agreement will be analyzed. The same samples as in the experiment of segmental labeling was used here again. Labeling was done by three labelers manually. Cohen's κ(kappa)[10] was used as the index of label agreement.

3.2.1 BI tier

Table 6 compares the accuracy of J_ToBI and X-JToBI labeling. Here, accuracy is defined as the ratio of labels that agreed with the labeling result of an expert labeler. Accuracy of X-JToBI labels were higher than J_ToBI labels nearly always.

As for inter-labeler agreement, κ is 0.64 and 0.73 in J_ToBI and X-JToBI respectively, but if we exclude the labels of disfluency and fillers from the calculation, there is no significant difference(κ is 0.69 and 0.71 in J_ToBI and X-JToBI). This suggests that newly proposed BI labels for disfluency and fillers has improved the inter-labeler reliability of prosodic labeling of spontaneous speech.

3.2.2 Tone tier

Table 7 shows accuracy of tone labels in BPMs. The accuracy of L% and L%H% was high, but that of L%HL% was noticeably low in the J_ToBI. On the other hand, the accuracy of L%HL% in X-JToBI was not very low.

Also, κ was 0.41 and 0.61 in J_ToBI and X-JToBI respectively as far as BPMs are concerned. Moreover, same tendencies were observed in "H-" and "H*+L".

Presumably, these improvements of the X-JToBI labeling were the byproducts of the very fact that it was a demanding labeling scheme. Because X-JToBI had more tone inventories than J_ToBI, and it required labels to be located at the exact locations of the corresponding physical (*i.e.* F0) events, lablers had to pay more attention for speech than in J_ToBI labeling which had fewer inventories and no strict requirement on the label location.

4 CONCLUSION

Performance of the segmental and prosodic labeling of spontaneous speech was evaluated using the sample data of CSJ. As for segmental labeling, it turned out that performance of automatic segmental labeling was lower than that of manual labeling, and, performance of spontaneous speech labeling was lower than that of read speech labeling. It seemed that the key to higher performance of automatic labeling of spontaneous speech consisted in the proper treatment of phonetic features of affricates and glide.

As for prosodic labeling, the new X-JToBI scheme showed higher accuracy and inter-labeler agreement compared to the traditional J_ToBI scheme. Introduction of new BI labels for filled-pauses and disfluency, on the one hand, and enlargement of BPM inventory seemed to be the main factor in the improvement.

Table 6: Accuracy of BI labels.
(numbers in () stands frequency of labels.)

J_ToBI			X-JToBI		
label	accuracy		label	accuracy	
1	91.3	(593)	1	94.9	(531)
2	74.0	(123)	2	74.4	(112)
3	70.5	(182)	3	75.1	(181)
2-	33.3	(1)	1+	0.0	(1)
1p	47.9	(16)	1+p	81.5	(9)
3-	–	(0)	2+	–	(0)
2m	80.5	(29)	2+b	66.7	(30)
2p	27.3	(11)	2+p	33.3	(8)
3m	0.0	(1)	2+pb	33.3	(4)
			D	83.3	(4)
			D+	44.4	(3)
—	—	—	<F	76.2	(7)
			F	91.0	(63)
			PB	33.3	(2)
Total	83.2	(956)	Total	86.1	(956)

Table 7: Accuracy of tone labels in BPMs.
(a)J_ToBI

	Correct label		
Labeling result	L%	L%H%	L%HL%
L%	144	15	11
L%+H%	7	57	6
L%+HL%	1	0	10
Accuracy[%]	94.7	79.2	37.0

(b)X-JToBI

	Correct label			
Labeling result	L%	L%H%	L%HL%	L%LH%
L%	360	34	17	0
L%+H%	12	157	3	0
L%+HL%	5	7	40	0
L%+LH%	0	0	0	0
Accuracy[%]	95.5	79.3	66.7	-

References

[1] K.Maekawa, H.Koiso, S.Furui, H.Isahara, "Spontaneous speech corpus of Japanese," Proc. 2nd International Conference on Language Resources and Evaluation, Athens, Greece, pp.947-952 (2000).

[2] C.W.Wightman, D.T.Talkn,"The aligner: Text-to-speech alignment using Markovmodels," In VanSanten et al., editors, *Progress in Speech Synthesis*, pp.313-323, Springer-Verlag (1996).

[3] T.Kawahara, et.al., "Free software toolkit for Japanese large vocabulary continuous speech recognition," Proc. of ICSLP 2000 (2000).

[4] T.Shinozaki, S.Furui, "A statistical analysis of individual differences in spontaneous speech recognition performance," Technical Report of IPSJ, 2001-SLP-39, pp.111-116 (2001).

[5] S. Young, et. al, *The HTK Book*, Entropic Research Lab (1999).

[6] Y.Sagisaka, K.Takeda, S.Katagiri, T.Umeda, H.Kuwabara, "A large-scale Japanese speech database," Proc. of ICSLP 1990, pp.1089-1092 (1990).

[7] A.Ljolje and M.D.Riley., "Automatic speech segmentation for concatenative inventory selection," Progress in Speech Synthesis, Springer, 305-311 (1997).

[8] Venditti, J., "Japanese ToBI Labelling Guidelines," Manuscript. Ohio State University, USA., 1995.

[9] K.Maekawa, H.Kikuchi, Y.Igarashi, J.Venditti, "X-JToBI: An extended J-ToBI for spontaneous speech," Proc. International Conference on Spoken Language Processing, Denver, U.S., pp.**-** (2002).

[10] Cohen, J., "A Coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, Vol.20, No.1, pp.37-46, 1960.