KOREAN PROSODIC BREAK INDEX LABELLING BY A NEW MIXED METHOD OF LDA AND VQ

Pyungsu Kang, Jiyoung Kang , Jinyoung Kim

Electronics Engineering of Chonnam National University
300 Yong-Bong Dong Puk-Ku Kwangju Korea
kpyung@dsp.chonnam.ac.kr, kimjin@dsp.chonnam.ac.kr

ABSTRACT

We present a new mixed method of LDA-VQ to predict Korean prosodic break index(PBI) for a given utterance. PBI can be used as an important cue of syntactic discontinuity in continuous speech recognition(CSR). Our proposed method, LDA-VQ model, consists of three steps. At the first step, PBI was predicted with the information of syllable and pause duration through the linear discriminant analysis(LDA) method. At the second step, syllable tone information was used to estimate PBI. In this step we used vector quantization(VQ) for coding the syllable tones and PBI is estimated by tri-tone model. In the last step, two PBI predictors were integrated by a weight factor. The LDA-VQ method was tested on 200 literal style spoken sentences. The experimental results showed 72% accuracy.

1. INTRODUCTION

An adequate prosody control is very important to achieve the naturalness of synthetic speech. There have been many research results on prosody analysis and synthesis. It's application researches in ASR are increasing, and some reported that prosody information was helpful in CSR or ambiguous sentence recognition[1][2]. Many researchers developed tone break index(ToBI) system using their own languages. ToBI is a standard for labeling prosodic information. ToBIs were used successfully in CSR, however, there is no ToBI standard for Korean, and Korean-ToBI is still under study. We introduced the PBI concept for enhancing the Korean CSR . PBI is an indicator representing how strongly successive words are linked in an utterance. For a given utterance, PBIs are determined by a listening test. That is, PBI is a psycho-acoustic index and PBI is related intimately to prosodic features in the word boundary of an utterance.

To estimate PBI, one must parameterize the prosodic discontinuity with such features as syllable duration, pause duration, syllable intensity and syllable tone. Until now there have been few studies done that predict Korean PBI.

Some researchers predicted PBI with syllable duration, pause or the mean value of syllable tone with multivariate regression method[3][4]. But did not utilize the above parameters integrally. Our experiment showed that the average of syllable tone is not a significant parameter in predicting Korean PBI.

Our research proposes another more efficient method, the LDA-VQ model. We predicted PBI with syllable and pause duration by the LDA method. A tri-tone concept was introduced to estimate PBI with pitch information. Tri-tone model is similar to the tri-gram language model or tri-phone model. At the word boundary, three boundary syllable tones are classified by VQ. Then the VQ index was used as the label of syllable tone, and has the same role with POS in tri-gram language model. The two predictors of LDA and VQ are tied as one predictor with a weighting factor to predict Korean PBI.

2. PREPARATION OF EXPERIMENTAL DATA

We constructed speech DB of 200 utterances, which were spoken by a female announcer. The constructed DB contained 2990 word boundaries. Three researchers performed listening test and the PBIs were graded with 4 levels, which was granted generally by most Korean prosody researchers. PBI levels are as follows

- 0 – Normal word boundary
- 1 – Boundary, making a minor grouping
- 2 – Intermediate phrase boundary
- 3 – Intonational phrase boundary

3. PREDICTION OF PBI WITH PHONEME AND PAUSE DURATION

Experimental research has proven that syllable and pause duration has a tendency to increase as PBI gets stronger. This is important parameter for predicting Korean PBI. As speaking style or speed can vary absolute syllable duration we introduced relative syllable duration, which is the duration ratios of boundary syllable and its neighboring syllables. Also pause duration must be adjusted with the speaking speed. We have not devised an appropriate method of pause normalization, thus pause normalization was not done. In the next sections the results of PBI prediction with duration information are shown.

3.1. Observation results on the Relative Syllable Duration

As discussed in the above, relative criterions are needed. In this paper we used two duration ratios and the duration ratio of penultimate and boundary syllables, the ratio of post and boundary syllables. Table 1 shows statistical characteristics of the duration ratios.

PBI\Param.	Statistics	PENUL	POST
0	Mean	0.898	1.094
	S.D	0.676	0.516
1	Mean	0.727	0.967
	S.D	0.324	0.386
2	Mean	0.609	0.849
	S.D	0.345	0.353
3	Mean	0.590	0.776
	S.D	0.300	0.348

Table 1: statistical characteristics of the duration ratios. PENUL is the ratio of penultimate and boundary syllable duration. POST is the ratio of post and boundary syllable duration.

From table 1 it can be concluded that the difference between boundary syllable duration and the neighbor syllable duration becomes longer as PBI increase. We performed t-test on the syllable ratios. The results showed that relative syllable duration is statistically significant in predicting PBI.

3.2. Observation Results on Pause Duration

Pause is a direct cue of the prosodic discontinuity. We can easily conjecture that pause duration becomes longer as PBI increases. Table 2 shows the average and standard deviation of pause versus PBI.

.

Class	PBI	Occurrence Number	mean	S.D
PBI	0	10	90	36
	1	18	123	72
	2	161	218	101
	3	132	386	151

Table 2: Pause duration versus PBI.

3.3. LDA Results with Duration Information

LDA is based on a discriminant function that is the linear combination of p variables that maximizes the distance between the group mean vectors. By finding the discriminant plane we can divide data into several group. Table 3 shows the LDA results with the syllable ratios and pause duration ratios.

Real \ Predict	0	1	2	3
0	53.68	48.35	0.47	0
1	75.0	61.26	4.5	0.45
2	6.0	38.37	37.55	53.0
3	0	3.01	15.79	81.0

Table 3: LDA results about the syllable duration ratios and pause duration (units: %).

4. PBI PREDICTION WITH PITCH

It is known that Korean has a LHLH pitch accent. LHLH means that any syllable pitch is relatively higher or lower than neighbor syllables in the mean sense, so it is too simple and abstract to model variant pitch patterns at prosodic boundaries. We need an adequate pitch model, which can reflect dynamic syllable tone variation, and predict the prosodic break index in the prosodic phrase. In this study we applied two approaches to model syllable tones. One is a parametric method and the other is a non-parametric method. In the sections below we describe the two proposed methods and show the experimental results.

4.1. Parametric modeling and its Application to PBI prediction

Assuming that the pitch pattern can be regarded as a slow varying signal, we were able to model a syllable pitch pattern with some smooth curves. In this paper the quadratic model is used.

$$p(t) = a_0 + a_1 t + a_2 t^2$$

Considering that the Korean pitch pattern is LHLH, we included the average pitch value as one of three parameters to determine the quadratic smooth curve. The first and second derivatives of syllable pitch pattern are used, for they reflect well the dynamics of syllable tones. So the adopted model parameters are as follows

- AP : mean value of pitch in a syllable
- ADP : the average of the first derivative of pitch in a syllable
- ADDP : the average of the second derivative of pitch in a syllable

The following table shows the rough pitch patterns (syllable tone) with AP, ADP and ADDP. From the table it is deduced that each parameter has its own physical meanings.

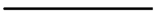
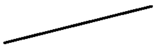
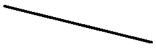
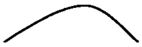
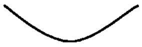
	ADP=0	ADDP≅0
	ADP>0	ADDP≅0
	ADP<0	ADDP≅0
	ADP≅0	ADDP<0
	ADP≅0	ADDP>0

Table 4: Pitch pattern and parameter values.

To predict PBI, the relative pitch pattern of boundary is important. We considered three pitch patterns: penultimate, boundary, and post syllables. With 3 modeling parameters and 3 syllable we applied 9 parameters to the LDA method. The following table shows the results.

Real \ Predict	0	1	2	3
0	20.71	20.9	29.33	29.05
1	20.0	22.79	28.37	28.34
2	20.0	38.78	38.78	23.67
3	18.0	19.55	27.82	34.59

Table 5: LDA results about the mean and derivatives per syllable considering penultimate, boundary, and post syllable (units: %).

From table 5 we can say that

- the mean value of pitch per syllable is not significant to predict PBI, and
- the quadratic model of syllable pitch doesn't reflect the dynamic features perfectly.

4.2. VQ Modeling of Syllable Pitch and its Application to PBI prediction

As discussed in the above parametric approach failed in PBI prediction and the reason is that parametric approach method can not model the dynamics of syllable tones. To overcome this problem we could increase the degree of the candidate curve. In this paper, however, we proposed another efficient modeling method. From the above experiments we concluded that syllable tones are dealt as one unit. So, we introduced the pattern recognition method(classifying method). In speech coding area, the approach is VQ method. By the way each syllable should have same dimension to apply VQ to the syllable tones. As syllables have different length so syllable tone(pitch pattern in syllable) has different vector dimension. It can be overcome by using linear interpolation. The VQ process is as follows.

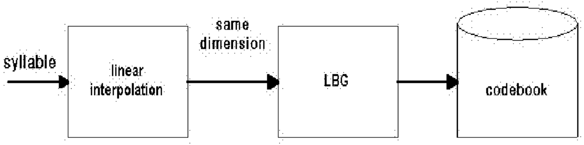


Figure 1: clustering process of syllable tones.

After processing, the VQ syllable tones are represented as a codebook indexes. We can regard the codebook index as a POS label in n-gram language model. Based on this idea, we proposed a tri-tone model, which predicts PBIs from a sequence of three consecutive syllable tone labels at word boundaries. Tones of penultimate, boundary and post syllables are considered. The PBI predictor is represented by the following equations.

$$p(BI \mid I_{penul}, I_{bound}, I_{post}) = \frac{N(BI \mid I_{penul}, I_{bound}, I_{post})}{\sum_{K} N(K \mid I_{penul}, I_{bound}, I_{post})}$$

$$PBI = \frac{\sum_{K} N(K \mid I_{penul}, I_{bound}, I_{post}) * BI}{\sum_{K} N(K \mid I_{penul}, I_{bound}, I_{post})}$$

, where N(K|I) is the number when index series is I , PBI is K and BI is the break index determined by listening test.

The codebook size affects the PBI prediction, and if the codebook sizes are big enough, prediction error can be reduced. But, considering the limited speech DB, we needed to determine their size and from some experiments we decided that the appropriate codebook size was 8.

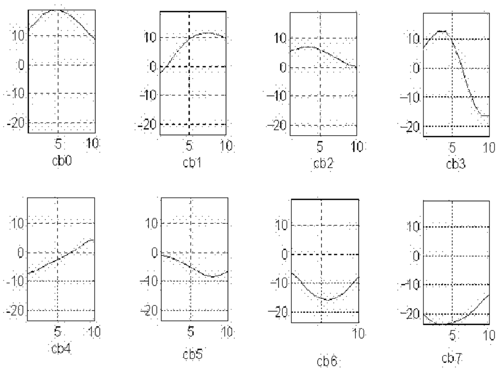


Figure 2 : Clustered codewords for syllable tone labeling.

5. LDA-VQ MIXED MODEL & RESULTS

In this section we describe the proposed integration method of the two PBI predictors proposed in the above. Figure 3 shows LDA-VQ model in which two predictors are tied using a weighting factor.

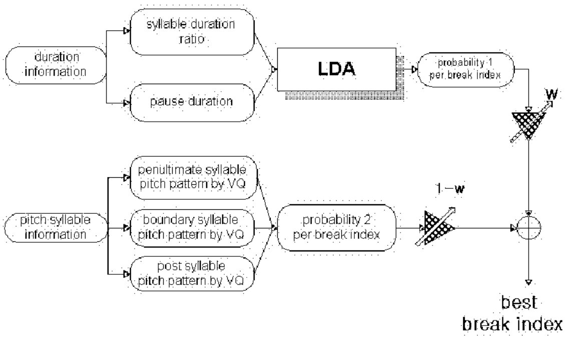

duration information
syllable duration ratio
pause duration
LDA
probability 1 per break index
w
pitch syllable information
penultimate syllable pitch pattern by VQ
boundary syllable pitch pattern by VQ
post syllable pitch pattern by VQ
probability 2 per break index
1−w
best break index

Figure 3: Proposed LDA-VQ model.

We call the mixed model LDA-VQ, as the LDA–based predictor is applied to the duration information and the VQ-based tri-tone model is applied to pitch information. We adopted a weighted sum mode because LDA and tri-tone have different measures. For obtaining optimal weight we predicted PBIs as weight value were varied. And we determined the best weight value for making the least error. Figure 4 shows classification rate versus weighting factor. The best weight value was 0.48. And table 6 shows optimal results. The average prediction error of 0.28 was obtained, this error is less than that of the LDA predictor with the value of 0.32.

6. CONCLUSION

In this paper a new model of LDA-VQ method was proposed to predict PBI with spoken sentences. The performances of the proposed method were evaluated with some prediction experiments. The experimental results showed that our model was effective to predict PBI. However, whether we can apply this model to many other speakers requires further studies. Our proposed method need to be tested to more utterances spoken by various speakers. On the other hand, energy information was not adopted to PBI predictions, and more research is needed on the effectiveness of energy information in PBI predictions. Our proposed method could be tested on real speech recognition systems.

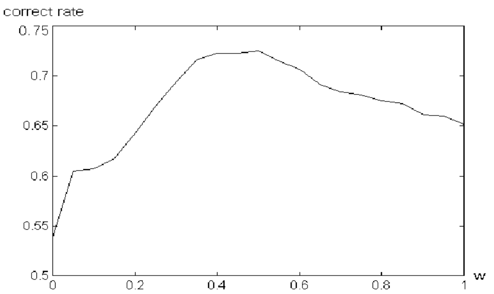


Figure 4: correction rates according to weights.

Real \ Predict	0	1	2	3
0	96.6	2.6	0.3	0.2
1	49.8	42.7	6.2	0
2	13.2	10.8	47.8	28.1
3	0.8	0.8	12.0	86.5

Table 6: LDA-VQ results about syllable duration ratio, pause duration and pitch pattern considering penultimate, boundary, and post syllable (unit : %).

7. REFERENCES

1. A.J. Hunt, A Generalized model for Utilizing Prosodic Information in Continuous Speech Recognition, Speech Technology Research Group, Univ. of Sydney, 1995
2. P. Taylor, A new Model of Intonation for Use with Speech Synthesis and Recognition,1992
3. ETRI, Spontaneous Speech Translation in Multimedia Environment, Ministry of Inf. And Comm. ROK, 1996.
4. KAIST AI, A Study of Korean Prosody and Discourse for the Development of Speech Synthesis/Recognition System, KT, 1997.
5. S.M. Kim, Rhythmic Units and Synthetic Structures in Korean, Ph.D thesis of Seoul Univ., 1997.
6. Y.J. Kim, S.H. Lee, Y.H. Oh, Relation between Prosodic Features and Dependency Relation, ICSP'97, 1997