Consistency Maintenance in Prosodic Labeling for Reliable Prediction of Prosodic Breaks

Youngim Jung
Dept. Knowledge Resources at Korea Institute of Science and Technology Information/
245 Daehang-no Yuseong-gu,
305-806 Daejeon, Republic of Korea
acorn@kisti.re.kr

Hyuk-Chul Kwon
Dept. Computer Science and Engineering at
Pusan National University/
San 30, Jangjeon-dong, Geumjeon-gu
Busan, 609-735, Republic of Korea
hckwon@pusan.ac.kr

Abstract

For the implementation of the prosody prediction model, large scale annotated speech corpora have been widely applied. Reliability among transcribers, however, was too low for successful learning of an automatic prosodic prediction. This paper reveals our observations on performance deterioration of the learning model due to inconsistent tagging of prosodic breaks in the established corpora. Then, we suggest a method for consistent prosodic labeling among multiple transcribers. As a result, we obtain a corpus with consistent annotation of prosodic breaks. The estimated pairwise agreement of annotation of the main corpus is between 0.7477 and 0.7916, and the value of K is between 0.7057 and 0.7569. Considering the estimated K, annotation of the main corpus has reliable consistency among multiple transcribers.

1 Introduction

The naturalness and comprehensibility of text-to-speech (TTS) synthesis systems are strongly affected by the accuracy of prosody prediction from text input. For the implementation of the prosody prediction model, large annotated speech corpora have been widely applied to both linguistic research and speech processing technologies as in (Syrdal and McGory, 2000). Since an increasing number of annotated speech corpora become available, a number of self-learning or probabilistic models for prosodic prediction have been suggested. To obtain reliable results from data-driven models, the corpus must be large scale, noise-free and annotated consistently. However, due to the limited range of tagged data with prosodic breaks that is used to learn or establish stochastic models at present, reliable results cannot be obtained. Thus, the reliability among transcribers was too low for successful learning of a prosodic model (Wightman and Ostendorf, 1994). In addition, the performance of ASR systems degrades significantly when training data are limited or noisy as in (Alwan, 2008).

In this study we propose a new methodology of training transcribers, annotating a corpus by multiple transcribers, and validating the reliability of intertranscriber agreement. This paper is organized as follows: we review related work on corpus annotation for speech and language processing tasks and method of measuring the reliability of consistency among multiple annotators in Section 2. Section 3 describes our observations on performance deterioration of the learning model due to inconsistent tagging of prosodic breaks in the established corpora. In Section 4, we suggest a procedure of constructing a medium-scale corpus, which are aimed at maintaining consistency in prosodic labeling among multiple annotators. Through a series of experiments during the training phase, the improvement of the agreement of multiple annotators is shown. The final experiment is performed in order to guarantee labeling agreement among five annotators. A brief summary and future work are presented in the final section.

2 Related Work

As linguistically-annotated corpora became critical resources, science of corpus annotation has been highlighted and evolved to reflect various interests in the field as shown in (Ide, 2007). In order to annotate linguistic information to large-scale corpora, two methods have been used; existing natural lan-

Proceedings of the Fifth Law Workshop (LAW V), pages 38–46,
Portland, Oregon, 23-24 June 2011.

guage processing (NLP) tools such as part-of-speech taggers, syntactic parsers, sentence boundary recognizers, named entity recognizers as have been used to generate annotations for ANC data (Ide and Suderman, 2006). Big advantages of using existing tools are that much cost and time can be saved and that the annotation result is consistent. In addition, it could obtain reliable accuracies and reduce the prohibitive cost of hand-validation by combining results of multiple NLP tools. However, tagging for all other linguistic phenomena is still mainly a manual effort as presented in (Eugenio, 2000). Thus, human annotators are required for tagging, correcting or validating the linguistic information although human annotators are very expensive and inconsistent in various aspects.

Linguists and language engineers have recognized the importance of the consistency of annotation among multiple annotators while they construct a large-scale corpus and have focused on how to measure the inter-annotator agreement. Their annotators had difficulties in discriminating one annotation category from others that are closely related to each other. Fellbaum et al. (1999) who performed a semantic annotation project which aimed at linking each content word in a text to a corresponding synset in WordNet found out that, with increasing polysemy, both inter-annotator and annotator-expert matches decreased significantly. As to measure the rate of agreement, Fellbaum et al. (1999) used a very simple measurement, the percentage of agreement in semantic annotation task. A greedy algorithm for increasing the inter-annotator agreement has been suggested by Ng et al. (1999). However, automatic correction of the manual tagging cannot reflect natural linguistic information tagged by human.

On the other hand, in prosodic annotation, the reliable measurement of intertranscriber agreement was studied by Beckman et al. (1994) initially, since the goal of the original ToBI system designers was to design a system with ‘reliability (agreement between different transcribers must be at least 80%)’, ‘coverage’, ‘learnability’, and ‘capability’. The designers and developers of adaptations of ToBI for other languages and dialects such as G-ToBI, GlaToBI and K-ToBI have proved the usability of their labeling system rather than have suggested the method of maintaining the intertranscriber agreement based on the aforementioned criteria (Grice et al., 1996; Mayo et al., 1996; Jun et al., 2000).

3 Problem Description

3.1 Obtaining a Large Scale Speech Annotated Corpus

In order to design and implement a prediction model of prosodic break, annotated corpus should be prepared. Recorded speech files and text scripts of Korean Broadcasting Station (KBS) News 9 were collected and manual annotation was conducted by two linguistic specialists. Each hand-labeled half of the selected script for prosodic breaks was cross-checked with the other half. The resultant corpus had 47,368 *eo-jeol*[1]s. The size of this corpus, however, does not seem to be sufficient. An easy way to construct a larger-scale corpus is using existing corpora in the field. To build a large volume of learning and testing data, annotated speech data from Postech speech groups were obtained. The Postech data included 122,025 *eo-jeol*s from Munhwa Broadcasting Corporation (MBC) news. Three types of break, viz., major breaks, minor breaks and no breaks, were annotated after each *eo-jeol* in KBS data (our initial data) and MBC data.

3.2 Performance Deterioration of Learning Models due to Inconsistent Annotation

KBS and MBC news data were selected, to examine the effect of prosodic breaks in corpora constructed by different groups on learning and testing. Only 46,526 *eo-jeol*s were randomly sampled from the MBC News corpus, whereas the entire KBS News data was used for learning and testing, to avoid potential side effects from the differing data size.

	KBS	**MBC (Postech data)**
Training Data	38,243	37,258
Testing Data	9,103	9,268

Table 1 Size of Training and Test data

[1] An *eo-jeol* in Korean can be composed up of one morpheme or several concatenated morphemes of different linguistic features which are equivalent to a phrase in English. This spacing unit is referred as an ‘*eo-jeol*’, ‘word’, or ‘morpheme cluster’ in Koeran linguistic literatures. We adopt ‘*eo-jeol*’ in order to refer to ‘an alphanumeric cluster of morphemes with a space on either side’.

C4.5 and CRFs were adapted in this experiment. The learning and testing was conducted in two phases. First, learning and testing of the prosodic break prediction models used a corpus constructed by a single group. Five-fold cross-validation was used for evaluating the models. Second, learning and evaluation of the models used a different corpus constructed by each group. The ratio of training to testing data (held-out data) was four to one. The results obtained from the first and second phases of learning and testing are presented in Table 2.

Algo-rithm	**1st Phase Precision (Learning -Testing)**		**2nd Phase Precision (Learning -Testing)**	
	KBS-KBS	MBC-MBC	KBS-MBC	MBC-KBS
C4.5	85.30%	62.53%	38.78%	44.96%
CRFs	84.65%	67.52%	37.96%	45.01%

Table 2 Experimental Results for Impact Analysis of Inconsistent Tagging

The prediction models performed well with C4.5 and CRFs learning algorithms when the model was trained and tested with KBS news data. However, its performance decreased drastically when the model was initially trained with KBS news data and subsequently tested with MBC news data. The performance of the learning model trained with MBC news data also deteriorated when tested with KBS data. These results suggest that serious performance deterioration is caused by data inconsistency rather than by the learning algorithm per se.

3.3 Analysis on Inconsistent Annotation

The deterioration of the performance presented in Section 3.2 is quite considerable, despite the fact that the same genre and level of prosodic break labeling system was selected. After analyzing the data, we identified three main reasons as follows.

(1) Perceptual Prominence of Prosodic Labeling Systems

Despite the fact that three types of prosodic break have been commonly used in the speech engineering field for a considerable time as shown in (Ostendorf and Veilleux, 1994), they have not been clearly defined or referenced in standard prosodic labeling conventions. In particular, the notion of the minor break is rather vague, whereas those of no break and major break are intuitively clear as in (Mayo et al., 1996). In the MBC news data labeled by Postech, sentences that had all prosodic breaks tagged as no break were frequently found, even if two long clauses exist in a sentence. Most sentences had been annotated only with no break. The speaking rate of news announcers on air is relatively fast and no obvious audible break seems to exist in their speech. However, Kim (1991) showed that even well-trained news announcers rarely read a sentence without breaks. Therefore, minor breaks need to be recognized not only by the duration of the break, but also by the tonal changes or lengthening of the final syllable as shown in (Kim, 1991; Jun, 2006; Jung et al., 2008).

(2) Different Perceptibility of Prosodic Breaks among Transcribers

Grice et al. (1996), Mayo et al. (1996) and Jun et al. (2000) have focused on reliability-agreement between different transcribers as the main criterion of evaluation. This fact indicates that individual labeling of a single utterance can differ, because each transcriber's recognition of the prosodic labeling system varies. And, the perceptibility of each transcriber differs. A large-scale corpus is necessary for modeling a data-driven framework, and the greater the number of transcribers cooperating, the poorer the intertranscriber agreement becomes. However, maintaining the intertranscriber agreements is often neglected as empirical work when researchers build and analyze a speech annotated corpus for implementation of the prosody model.

(3) Syntactic or Semantic Ambiguities

A single sentence with syntactic ambiguities has several different interpretations. In spoken language, prosody prevents garden path sentences and enables resolution of syntactic ambiguity as shown in (Kjelgaard and Speer, 1999; Schafer, 1997).

Sentences such as the one in the following example (E1) can be grammatically constructed with multiple syntactic structures[2].

(E1) 고속버스가 중앙선을 침범해 마주오던 승용차를 들이받았습니다.

a. *Gosogbeoseuga // jung-angseon-eul # chimbeom-hae /// maju-odeon # seung-yongchaleul // deul-ibad-ass-seubnida*

'An express bus drove over the center line and

[2] In examples, letters in italics denote phonetic transliteration of Korean; hyphens in transliteration are used for segmentation of syllables.

rammed into an oncoming car.'

b. *Gosogbeoseuga /// jung-angseon-eul # chimbeomhae // maju-odeon # seung-yongchaleul /// deul-ibad-ass-seubnida*
'An express bus rammed into an oncoming car which drove over the center line.'
#: no break, //: minor break, ///: major break

The prosodic phrasing in both (a) or (b) can be correct, depending on the sentence's syntactic structure. The pattern in (E1) is quite frequent in Korean, particularly in situations where the topic is broad. This kind of syntactic ambiguity needs to be resolved by semantic or pragmatic information, since it cannot be resolved using syntactic information only.

As we previously mentioned, three main problems arise when annotated speech data are both constructed by multiple labelers in a research group and the data are collected from different groups. Considering the impact of the quality of annotated corpora on the data-driven models, the overall procedure of corpus construction including the data collection and preprocess, labeling system selection and intertranscriber agreement maintenance should be designed and then evaluated as shown in Section 4.

4 Corpus Building

4.1 Selection of Prosodic Labeling System

In this paper, we define seven types of prosodic break in combination with phrasal boundary tones since a prosodic break cannot be separated from a boundary tone. Our seven types are defined as follows:

(1) **Major break with falling tone**: For cases with a strong phrasal disjuncture and a strong subjective sense of pause. The positions of major breaks generally correspond to the boundaries of intonational phrases (marked '///L').
(2) **Major break with rising tone**: For cases with a strong phrasal disjuncture but a weak subjective sense of pause length (marked '///H').
(3) **Major break with middle tone**: In real data, major breaks with middle tone (or major breaks without tonal change) are observed as in (Lee, 2004), although they have no definition or explanation in K-ToBI. They have been observed in very fast speech such as headline news utterances (marked '///M').
(4) **Minor break with rising tone**: For cases with a minimal phrasal disjuncture and no strong subjective sense of pause. The positions of minor breaks correspond to the boundaries of accentual phrases with rising tone. When an utterance is so fast that a pause cannot be recognized clearly, minor breaks are realized by tonal changes or segment lengthening of the final syllable (marked '//H').
(5) **Minor break with middle tone**: For cases with prosodic words in compound words, such as compound nouns or compound verbs. Breaks between noun groups in a compound word or between verbs in a compound verb may be realized when the overall length of a compound word is long, whereas a break is absent in a short compound word (marked '//M').
(6) **Minor break with falling tone**: For cases with minimal phrasal disjuncture and no strong subjective sense of pause. The positions of minor breaks correspond to the boundaries of accentual phrases with falling tone.
(7) **No break**: For internal phrase word boundaries. There is no prosodic break between one-word modifiers and their one-word partners or between a word-level argument and its predicate, because the two words are syntactically and semantically combined (marked '#').

The seven types of prosodic break are mapped to K-ToBI break indices, enabling further reusability of the corpus labeled by the suggested break types.

K-ToBI		Suggested Prosodic Breaks
Break Index	0	No Break (#)
	1	Minor Break (//L)
	2	Minor Break (//H, //M)
	3	Major Break (///H, ///M, ///L)
Tone Index	Ha, H%	H
	La, L%	L
	L+	M

Table 3 Mapping between break indices of K-ToBI and the suggested prosodic breaks

Jun et al. (2000) showed that the tonal pattern agreement for each word was approximately 36%

for all labelers and this low level of agreement appears to be due to the nature of the tonal pattern. Although fourteen possible AP (Accent Phrase) tonal patterns exist, these variations are neither meaningful nor phonologically correct. We concluded that the final phrasal tones are sufficient for the recognition of prosodic boundaries.

4.2 Data Selection and Preprocessing

In this study, KBS news scripts (issued January, 2005 ~ June, 2006) were collected as a raw corpus from web. Although the speech rate of TV news speech is faster than that of general read speech, announcers are trained to speak Standard Korean Language and to generate standard pronunciations, tones and breaks. In addition, individual stylistic variation is restricted in the announcer's speech.

The text formats of news scripts extracted from the web are unified. Then, sentences or expressions in news scripts differing from those in real sentences in multimedia files are revised according to the real utterances of the announcer. The selection and revision of the sentences is performed according to the following criteria.

1) Headline news sentences uttered by one female announcer are collected.
2) Minimum of five *eo-jeol*s are included in one sentence.
3) Real speech of news script read by the announcer is considered as primary source of prosodic break tagging for transcribers.
4) Sentences in the news script are deleted unless they are read by the announcer in real speech files.
5) Between 1-3 *eo-jeol*s in news scripts differing from those in speech files are revised according to the real speech if there is no semantic change.
6) Sentences in the news script differing considerably from those in speech files are deleted.
7) Words or phrases in the news script differing from those in speech files due to spelling/grammar errors are not corrected manually. They are corrected automatically by the PNU grammar checker, which shows over 95% accuracy as in (Kwon et al., 2004).

4.3 Training Transcribers

The most reliable method of maintaining the consistency and accuracy of prosodic breaks by multiple transcribers is for each well-trained transcriber to annotate prosodic breaks in the entire corpus. Then the majority of the tagging results among multiple transcribers are selected as an answer for the target *eo-jeol*. However, this method where all transcribers annotate the same corpus in depth is too time consuming and costly. Due to time and cost constraints, most related studies use a simpler method. If the size of the corpus is small, then a professional linguist annotates the entire corpus as in (Maragoudakis et al., 2003). If the size of corpus is large, more than two transcribers divide the corpus by the number of transcribers and each transcriber annotates his/her own part as in (Wightman and Ostendorf, 1994; Viana et al., 2003). Unless the transcribers are trained and the reliability of the intertranscriber agreement is validated, consistency of annotation by multiple transcribers cannot be assured. Hence, a method for maintaining the reliability of the intertranscriber agreement of prosodic breaks is suggested in this paper.

The overall procedure of training the transcribers, annotating the main corpus with prosodic breaks and validating the reliability of tagging consistency among multiple transcribers is illustrated in Figure 1.

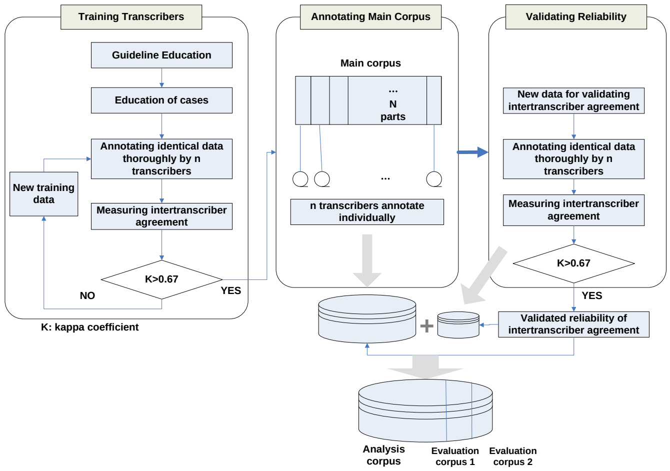
Training Transcribers
Guideline Education
Education of cases
Annotating identical data thoroughly by n transcribers
New training data
Measuring intertranscriber agreement
K>0.67
NO
YES
K: kappa coefficient
Annotating Main Corpus
Main corpus
...
N
parts
...
n transcribers annotate individually
Validating Reliability
New data for validating intertranscriber agreement
Annotating identical data thoroughly by n transcribers
Measuring intertranscriber agreement
K>0.67
YES
Validated reliability of intertranscriber agreement
Analysis corpus
Evaluation corpus 1
Evaluation corpus 2

Figure 1 Overall Procedure of Corpus Building

Firstly, guidelines are provided for transcribers to familiarize themselves with the prosodic labeling system suggested in Section 4.1. Secondly, in order to improve the awareness of the length or strength of each prosodic break type in detail, transcribers repeatedly listen to speech files corresponding to several paragraphs in news scripts. In addition, WaveSurfer Version.1.8.5, which is an open source program for visualizing and manipulating speech, is utilized for transcribers to examine the pitch contour, waveform, and power plot of speech files.

In the training phase, five transcribers annotate the same data with prosodic breaks at the same time and then compare the results of their annotations, and discuss and repeatedly correct the various errors until reliable agreement among them is reached. The data used for this intertranscriber agreement training is given in Table 4.

	1st	2nd	3rd	4th
# *eo-jeol*s	422	544	491	711
# sentences	35	49	42	32

Table 4 Data used in intertranscribers training

After mastering the guidelines and training with each data set, specific reasons for inconsistency among transcribers were analyzed and their solutions were educated as follows:

(1) Prosodic breaks were inserted due to announcers' emphasis on a certain *eo-jeol*, mistakes in reading the sentence or the habit of slowing down two or three *eo-jeol*s from the end of a sentence. Some transcribers recognized these as speakers' errors and corrected them in their annotations. On the other hand, others annotated prosodic breaks according to what they heard, regardless of errors. Due to these differing policies on annotation, the resultant annotation of prosodic breaks among transcribers is not consistent, as shown in example (E2).

(E2) 더욱 심각해지고 (///H, #)[3] 있습니다.
deo-ug simgaghaejigo iss-seubnida.
more serious become progress +EM[4]
"(sth) becomes more serious"

Inconsistency derived from these speakers' errors should be deleted.

(2) If the speech rate of the announcer is too fast for some transcribers to perceive audible breaks between two *eo-jeol*s, they omitted the minor break, whereas others put a minor break in the same place, as shown in (E3).

(E3) 그러나 (#, **//L**) 질병관리본부는
geuleona jilbyeonggwanlibonbu-neun
however Korea Center for Disease Control+TP and Prevention+TP
"However, the Korea Center for Disease Control and Prevention"

In this case, transcribers need to pay attention to whether the final tone of the target *eo-jeol* is rising or falling. In order to reduce inconsistency derived from missing breaks, transcribers repeatedly practice while listening to similar patterns.

(3) If only one annotator selects a different type of prosodic break than the others for the answer of the same place, he/she must change his approach in annotating prosodic breaks.

(4) Wightman and Ostendorf (1994) and Ross and Ostendorf (1996) have revealed that there is prosodic variability even for news speech data. The announcer showed variability in the location, strength or length, and tonal change in our news data as well. For example, the announcer occasionally put a minor break between two *eo-jeol*s consisting of a time expression, as shown in (E4).

(E4) a. 지난 //H 2002 년 오늘,
jinan //H 2002nyeon oneul,
past 2002year this day
"(on) this day 2002,"

b. 지난 # 2000 년 1 월
jinan # 2000nyeon 1wol
past 2000year January
"(in) January 2000,"

For a time expression including less than four *eo-jeol*s, no break should be marked in it.

Discussion and education such cases described above after annotating new training data sets repeats till the intertranscriber agreement is sufficiently high. The intertranscriber agreement in annotating seven-level prosodic breaks including tonal changes is shown in Table 5.

Agreement	**Cumulative rate (%)**

[3] The correct answer among different annotations is underlined.

[4] Notes on abbreviations of Korean grammatical morphemes are as follows: EM for ending markers, TP for topical postposition, LCM for locative case marker, OCM for objective case marker, PEC for pre-ending denoting continuous

	1st	2nd	3rd	4th
Five (all) agreed	43.84	50.55	55.80	57.67
At least four agreed	60.90	68.20	73.52	75.53
At least three agreed	81.75	87.50	90.84	91.70

Table 5 Intertranscriber agreement in training

The cumulative rate of agreement of more than half of the transcribers (n+1/2) is measured by approximate figures. Specifically, the rate of the intertranscriber agreement is calculated with the cumulative rate at which all five transcribers agreed, at least four of them agreed, and at least three of them agreed. The resultant agreement of the first experiment is quite low, though the first experiment was performed after the transcribers had familiarized themselves with the guidelines and studied many examples. The intertranscriber agreement in annotating data with seven-level prosodic breaks increases continuously with repeated training and experiments. This indicates that educating transcribers with guidelines and examples is not sufficient, and training of transcribers is required prior to annotation of the main corpus with specified tagging classes by multiple transcribers.

In order to review how accurately each individual transcriber annotates the corpus, the annotation accuracy of each individual transcriber is estimated. The prosodic break type for which at least three of them agreed is considered as the answer. The annotation result of each transcriber is compared to the answer, and then the accuracy is estimated by counting the number of annotations that match the answers. Table 6 shows the estimated annotation accuracy of five transcribers from the 1st to the 4th experiment.

Transcriber	Estimated accuracy (%)			
	1st	2nd	3rd	4th
A	94.51	84.00	86.32	91.56
B	78.03	85.26	89.24	93.25
C	78.03	93.05	94.39	94.02
D	88.44	90.32	90.36	90.64
E	82.37	83.79	84.08	89.11

Table 6 Estimated accuracy of each transcriber

Although there are individual variations, the estimated accuracy of the transcribers increases steadily.

After the four experiments, the cumulative rate of agreement of more than half of the transcribers reached 91.70% and the estimated accuracy of individual transcribers increased to 89.11~94.02%. Hence, an objective and reliable measurement for intertranscriber agreement is required in order to decide whether the training is sufficient.

The most commonly used methods to assess the level of agreement among transcribers are pairwise analysis and Kappa statistics. The reliability of intertranscriber agreement of the four experiments has been assessed with these two measurements and the result is given in Table 7.

Measurement	1st	2nd	3rd	4th
Pairwise analysis	0.6385	0.6969	0.7375	0.7477
Kappa statistics	0.5783	0.6464	0.6938	0.7057

Table 7 Reliability of intertranscriber agreement

Since the value of K is greater than 0.67 in the 3rd and 4th experiment, the intertranscriber agreement for annotating prosodic breaks is considered to have reached a reliable level as shown in (Carletta, 1996). Then annotation of the main corpus is performed.

The main corpus comprising 29,686 *eo-jeol*s is divided into five parts. Each partition is assigned to the trained five transcribers and annotation is independently performed. WaveSurfer, which is used in the training phase, is also used in the annotation phase for the display and annotation of speech. Transcribers may openly discuss their annotations, even though they annotated different parts of the main corpus.

4.4 Validation of Reliability of Intertranscriber Agreement

Since each individual transcriber annotated a different part of the main corpus, the reliability of intertranscriber agreement cannot be measured directly. We assume that intranscriber agreement does not change dramatically before and after annotation of the main corpus.

Hence, another data set including 1,149 *eo-jeols* (46 sentences), with a size 1.5x larger than that of the data set used in the 4th experiment, is collected and used instead, in order to validate the reliability of agreement. Immediately after annotation of the main corpus, the final experiment is performed following the procedure performed in the training phase, except for the education steps. The five transcribers annotated the same data in depth, however, they worked independently. They were not allowed to discuss prosodic labeling. Pairwise analysis and Kappa statistics are used in measuring intertranscriber agreement on the validation data set. The pairwise agreement and K found in the validation experiment after annotation of the main corpus was 0.79 and 0.76, respectively.

Both agreement figures are greater than those found in the prior experiments, which were repeated four times in the training phase. Based on this result, annotation of the main corpus is also considered to be part of training of transcribers.

According to our assumption, the estimated intertranscriber agreement of annotation of the main corpus annotation is between the agreement of the prior and post experiments, as shown in Figure 2.

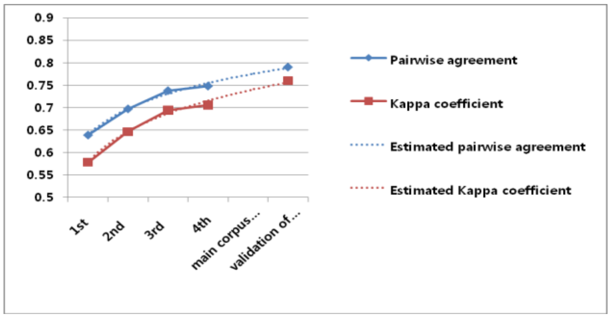

Figure 2 Estimated intertranscriber reliability in annotation of main corpus

The estimated pairwise agreement of annotation of the main corpus is between 0.7477 and 0.7916, and the value of K is between 0.7057 and 0.7569. Considering the estimated K, annotation of the main corpus has reliable consistency among multiple transcribers.

As a result, we obtained a corpus with consistent annotation of prosodic breaks. The data used in validation experiment is included as well. The statistics of the constructed corpus is shown in Table 8.

Data	**# *eo-jeols***	**# sentences**
Data set from validation experiment	1,149	46
Main corpus	29,663	1,319
Total	30,812	1,365

Table 8 Size of resultant corpus

It took approximately three months for us to train transcribers, annotate main corpus and validate the reliability of intertranscriber agreement in the main corpus. Considering the size of the constructed corpus, three months might be regarded as a considerable amount of time for researchers who want to build a large-scale annotated corpus. However, most time was spent on analyzing the inconsistencies among transcribers in initial experiments during the training step. Hence, if transcribers are trained following the suggested method in this paper, the amount of time for transcribers to annotate the target corpus with reliable consistency will decrease dramatically compared with the time for all transcribers to annotate prosodic breaks in the entire corpus.

5 Conclusions

In this study, potential problems in the construction, collection and utilization of a speech annotation corpus have been identified, and a solution for each type of problem has been suggested. The overall procedure of training transcribers, tagging the main corpus and validating the reliability of intertranscriber agreement on the main corpus has also been specifically described. As a result, we obtained a corpus with consistent annotation of prosodic breaks. The estimated pairwise agreement of annotation of the main corpus is between 0.7477 and 0.7916 and K is between 0.7057 and 0.7569. The suggested method for constructing a consistently annotated corpus and validating the consistency of the resultant annotation must be applied prior to implementation of data-driven models for predicting prosodic breaks. As our future work, the resultant corpus will be used for building a robust prediction model of prosodic boundary.

In addition, the method can be utilized for semantic annotation tasks, discourse tagging and others, which have a similar problem due to the differing perceptions of transcribers in recognizing the closely related categories.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (2010-0028784).

References

Abeer Alwan. 2008. Dealing with Limited and Noisy Data in ASR: a Hybrid Knowledge-based and Statistical Approach, Proc. Interspeech 2008, Brisbane Australia, , pp. 11-15.

Amy J. Schafer. 1997. Prosodic Parsing: The Role of Prosody in Sentence Comprehension, University of Massachusetts.

Ann K. Syrdal and Julia McGory. 2000. Inter-transcriber Reliability of ToBI Prosodic Labeling, Proc.Interspeech 2000, pp. 235-238.

Barbara Di Eugenio. 2000. On the usage of Kappa to evaluate agreement on coding tasks, Proc. Second International Conference on Language Resources and Evaluation, pp.441-444.

Catherine Mayo, Matthew Aylett, D. Robert Ladd. 1996. Prosodic Transcription of Glasgow English: An Evaluation Study of GlaToBI, Proc. ESCA Workshop on Intonation: Theory, Models and Applications, Athens Greece, pp.231-234.

Christiane Fellbaum, Joachim Grabowski and Shari Landes. 1999. Performance and Confidence in a Semantic Annotation Task, WordNet: An Electronic Lexical Database etd. Fellbaum, MIT Press, London.

Colin W. Wightman and Mari Ostendorf. 1994. Automatic Labeling of Prosodic Patterns, IEEE Transactions on Speech and Audio Processing, 2(4):469-481.

Hee Tou Ng, Chung Yong Lim and Shou King Foo. 1999. A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation, Proc. ACL SIGLEX Workshop on Standardizing Lexical Resources pp. 9-13.

Ho-Young Lee. 2004. H and L are Not Enough in Intonational Phonology, Korean Journal of Linguistics, 39:71-79.

Hyuk-Chul Kwon, Mi-young Kang and Sung-Ja Choi. 2004. Stochastic Korean Word Spacing with Smoothing Using Korean Spelling Checker, Computer Processing of Oriental Languages, 17:239-252.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic, Computational Linguistics, 22(2):249-254.

K. Ross and M. Ostendorf. 1996. Prediction of abstract prosodic labels for speech synthesis, Computer Speech and Language, 10(3):155-185.

M. Céu Viana, Luís C. Oliveira and Ana I. Mata. 2003. Prosodic Phrasing: Machine and Human Evaluation, International Journal of Speech Technology, 6:83-94.

M. Maragoudakis, P. Zervas, N. Fakotakis and G. Kokkinakis. 2003. A Data-Driven Framework for Intonational Phrase Break Prediction, Lecture Notes in Computer Science, 2807: 189-197.

M. Ostendorf and N. Veilleux. 1994. A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location, Computational Linguistics, 20(1):27-54.

Margaret M. Kjelgaard and Shari R. Speer. 1999. Prosodic Facilitation and Interference in the Resolution of Temporary Syntactic Closure Ambiguity, Journal of Memory and Language, 40:153-194.

Martine Grice, Matthias Reyelt, Ralf Benzmuller, Jörg Mayer and Anton Batliner. 1996. Consistency in Transcription and Labelling of German Intonation with GToBI, Proc. Interspeech1996, pp. 1716-1719.

Mary E. Beckman, John F. Pitrelli and Julia Hirschberg. 1994. Evaluation of Prosodic Transcription Labeling Reliability in the ToBI Framework, Proc. Interspeech 1994, pp. 123-126.

Nancy Ide. 2007. Annotation Science From theory to Practice and Use: Data Structures for Linguistics Resources and Applications, Proc. Bienniel GLDV Conference, Tübingen, Germany.

Nancy Ide and Keith Suderman. 2006. Integrating Linguistic Resources: The American National Corpus Model, *Proceedings of the Fifth Language Resources and Evaluation Conference*, Genoa, Italy.

Sangjun Kim. 1991. Study on Broadcast Language, Hongwon, Seoul.

Sun-Ah Jun. 2006. Prosody in Sentence Processing: Korean vs. English, UCLA Working Papers in Phonetics, 104:26-45.

Sun-Ah Jun, Sook-Hyang Lee, Keeho Kim, Yong-Ju Lee. 2000. Labler agreement in Transcribing Korean Intonation with K-ToBI, Proc. Interspeech 2000, pp. 211-214.

Youngim Jung, Sunho Cho, Aesun Yoon and Hyuk-Chul Kwon. 2008. Prediction of Prosodic Break Using Syntactic Relations and Prosodic Features, Korean Journal of Cognitive Science, 19(1):89 -105.

WaveSurfer. WaveSurfer ver.1.8.5, http://crfpp.sourceforge.net/.