electronics

MDPI

Article

Korean Prosody Phrase Boundary Prediction Model for Speech Synthesis Service in Smart Healthcare

Minho Kim [1], Youngim Jung [2,3] and Hyuk-Chul Kwon [4,*]

[1] Department of Software, Catholic University of Pusan, Busan 46252, Korea; minho@cup.ac.kr
[2] KESLI Secretariat, Korea Institute of Science and Technology Information, Seoul 02456, Korea; acorn@kisti.re.kr
[3] KISTI Campus, University of Science and Technology, Seoul 02456, Korea
[4] School of Computer Science and Engineering, Pusan National University, Busan 46241, Korea
* Correspondence: hckwon@pusan.ac.kr

Abstract: Speech processing technology has great potential in the medical field to provide beneficial solutions for both patients and doctors. Speech interfaces, represented by speech synthesis and speech recognition, can be used to transcribe medical documents, control medical devices, correct speech and hearing impairments, and assist the visually impaired. However, it is essential to predict prosody phrase boundaries for accurate natural speech synthesis. This study proposes a method to build a reliable learning corpus to train prosody boundary prediction models based on deep learning. In addition, we offer a way to generate a rule-based model that can predict the prosody boundary from the constructed corpus and use the result to train a deep learning-based model. As a result, we have built a coherent corpus, even though many workers have participated in its development. The estimated pairwise agreement of corpus annotations is between 0.7477 and 0.7916 and kappa coefficient (*K*) between 0.7057 and 0.7569. In addition, the deep learning-based model based on the rules obtained from the corpus showed a prediction accuracy of 78.57% for the three-level prosody phrase boundary, 87.33% for the two-level prosody phrase boundary.

Keywords: Korean prosody phrase boundary prediction; speech synthesis; text-to-speech (TTS); rule-based system; bidirectional LSTM-CRF

Citation: Kim, M.; Jung, Y.; Kwon, H.-C. Korean Prosody Phrase Boundary Prediction Model for Speech Synthesis Service in Smart Healthcare. *Electronics* **2021**, *10*, 2371. https://doi.org/10.3390/electronics10192371

Academic Editor: Yeong-Seok Seo

Received: 27 August 2021
Accepted: 24 September 2021
Published: 28 September 2021

1. Introduction

Speech processing technology has demonstrated great potential to provide beneficial solutions for both patients and doctors in smart healthcare. Recent advances in speech processing technology and other advanced technologies, including the Internet of Things (IoT) and communication systems, have significantly advanced contemporary healthcare systems [1–3]. In particular, recent innovations in deep learning, the advent of IoT and new communication systems have opened up various possibilities for medical systems. The voice interface represented by speech synthesis and speech recognition can be used to transcribe medical documents, control medical devices, mitigate speech and hearing impairments, and support the visually impaired. In addition, it can be used as a biomarker in diagnosing psychological disorders.

- Speech interfaces for doctors and patients are increasingly being developed and implemented in clinical practice. Speech technology is an essential means of reducing the cost of traditional medical records in healthcare systems [4,5]. In a previous study [6], researchers found that speech recognition of clinical documents saves time, increases efficiency, and enables more detailed notes with relevant details. Moreover, speech-technology-based interfaces can support patient's overall hospital experiences. In particular, environmental control assistance (e.g., device control, audio level control, nursing assistance requests, decision-making assistance) can aid in the recovery of patients with reduced mobility [7].

- People in developing countries typically lack literacy skills (i.e., reading and writing). People with low literacy are one third more likely to misunderstand prescribed medications, particularly due to the terminology used in the medical field [8]. Text-based health care is not very useful for the illiterate, the blind, or those without computer skills. Phonetic language is a plausible interaction method for the illiterate, and speech-based medicine may be ideal for residents of developing countries.
- Language skills can assist individuals with hearing problems or speech, language, or language impairments in order to communicate effectively [9]. Speech recognition plays a vital role in speech therapy applications that require the recognition of user utterances [10]. Similarly, speech synthesis can teach a user how to pronounce a word or sentence to reinforce correct pronunciation in speech therapy activities. Thus, speech synthesis and speech recognition systems can be utilized in therapy to improve the quality of human communication [11,12].
- Recent studies have shown the possibility of using spoken language as an effective biomarker for diagnosing psychological disorders. State-of-the-art deep learning models have used language to improve performance in emotion recognition, depression, anxiety, stress, pain, and suicidal behavior detection [13–15]. Deep learning models play an essential role in modeling and diagnosing various psychological disorders using speech signals.

This study focuses on a method for synthesizing natural speech when using speech technology in the medical field. Speech synthesis, also known as text-to-speech (TTS), is an important technology that aims to convert text to speech. The comprehensibility and accuracy of text-to-speech (TTS) synthesis systems are strongly affected by accurate prosody prediction from text input and by the audible realization of prosody in synthetic speech output. In particular, the performance of fundamental frequency contour generation, duration, and pause insertion modules is heavily dependent on the ability of the prosodic phrase break component to place boundaries at appropriate points [16].

An annotative speech corpus has been widely applied in language research and speech processing techniques to predict prosody phrase breaks automatically [17]. As more annotated speech corpora become available, several self-learning or probabilistic models for prosodic prediction have been suggested. These include hidden Markov models, classification and regression trees (CART) [18,19], transformational rule-based learning (TRBL) [20], and Bayesian networks. These models have been used to predict prosodic phrasal boundaries in English, Spanish, Korean, and Greek. The annotated speech data must be sufficiently large and in agreement with different transcribers that fall within a reliable scope to obtain reliable results from these data-driven models. To date, however, the reliability of automated transcribers has been insufficient for the successful operation of an automatic prosody recognizer [21].

In this study, we investigate applications of a rule-based natural language processing method and a linguistic knowledge base in the area of speech. The application of manually constructed rules created by researchers thus far has been criticized for their considerable expense, lack of systematic patterns in the data, and insufficient applicability to other domains [20]. Because only a tiny range of tagged data with prosodic breaks is currently used to learn or establish stochastic models, reliable results cannot be obtained. Additionally, various inaccuracies such as spelling and grammar errors present in the original corpus and inconsistent tagging have often led to meaningless results. This study proposes a new methodology for the reliable prediction of prosodic breaks using linguistic knowledge and bi-gram information obtained from a small-scale corpus. Contrary to many computational tasks whose answers are fixed, multiple answers can be acceptable in predicting optional prosodic breaks; thus, different methods need to be adopted to solve the problem.

We begin by discussing the three steps in our approach, which include the following. (1) A method is utilized to maintain consistency in prosodic labeling between different transcribers. The method proposed adopts a simple but more appropriate prosody labeling system and training procedure for such labelers. (2) The predicted locations of mandatory

prosodic breaks are processed with partial parsing analysis of syntactic structures. Based on this, rules are established that predict mandatory prosodic breaks. (3) The proposed model is evaluated using various performance measures adopted in previous studies for comparison.

2. Related Works

Approaches to prosody phrase boundary prediction are divided mainly into rule-based approaches and statistical approaches. The former is a method of making and using rules for predicting the prosody phrase boundary from linguistic information. The latter is a method of constructing and using a statistical model suitable for predicting the prosody phrase boundary using statistical values derived from the corpus.

The prosody phrase boundary prediction rules are created either entirely manually by experts, or automatically or semi-automatically using a corpus. Hand-made rules are independent of specific language resources and have high rule accuracy. However, since a lot of time and effort is required to establish rules, it is not easy to handle any exceptions occurring in actual living language. On the other hand, an automatically created rule has the advantage of being easy to construct and has a wide application range. However, it has the problem of being dependent on a specific corpus.

The statistical approach has the advantage of building a statistical model with good scalability, a wide application range, and relatively high overall accuracy by using a large-capacity prosody boundary analysis corpus. Artificial neural networks based on dense vector representations have shown excellent performance in predicting prosody phrase boundaries in recent years [14,22–30]. This is because high-quality word embeddings and effective deep learning techniques have been developed. In particular, deep learning techniques enable multi-level automatic feature representation learning. However, since it is necessary to construct a reliable analysis corpus of a specific size or level of prominence to extract meaningful features, much time and effort are required to construct the corpus. In particular, as the prosody phrase boundary analysis relies heavily on the subjective judgment of the commentator, more effort is required to maintain the consistency of the annotation.

We analyzed two annotation corpora constructed in previous studies to analyze the reliability of the annotation corpus for predicting the boundary of the Korean prosody phrase. ETRI speech data consisted of 18,253 words extracted from news scripts. POSTECH data contained 122,025 words from MBC News and was automatically transcribed [31]. In addition, recorded voice files and text scripts of the KBS News 9 program were additionally collected, and manual annotations were performed by two language experts. These three corpora are represented by three types of prosody boundaries, including major breaks, minor breaks, and no breaks after each word. The three types of prosody phrase distributions are shown in Table 1.

Table 1. Distribution of the three types of prosodic break.

	POSTECH Data		**ETRI Speech Data**		**KBS News Data**	
	Quantity	**(%)**	**Quantity**	**(%)**	**Quantity**	**(%)**
Major Break	23,829	19.53	4573	25.05	10,059	21.24
Minor Break	4479	3.67	3844	21.06	14,272	30.13
No Break	93,717	76.80	9836	53.89	23,037	48.63
Total	122,025	100	18,253	100	47,368	100

Table 1 shows the percentage distribution of the three prosodic breaks in each corpus obtained from the different sources. Even though the same genre and prosodic break labeling system level were selected, the difference was quite considerable. After analyzing the collected data, we identified four main reasons for this.

Although three types of prosodic break have been commonly used in the speech engineering field for a considerable time [32], they have not been clearly defined or referenced in standard prosodic labeling conventions. In particular, the notion of a minor break is rather vague, whereas no breaks and major breaks are intuitively clear [33]. In the POSTECH data, sentences with all prosodic breaks tagged as no breaks were frequently found, as shown in Box A1. The above sentence had been annotated only with no break because of the lack of the ability to distinguish minor breaks. The speaking rate of news announcers on air is relatively fast, and there are no obvious audible breaks in their speech. However, even well-trained news announcers rarely read sentences without breaks. Therefore, minor breaks need to be recognized not only by the duration of the break but also by tonal changes or lengthening of the final syllable [34].

The original ToBI (Tone and Break Indices) system designers aimed to develop a system of intonational transcription with improved reliability, coverage, learnability, and additional capabilities. Most ToBI-like system designers who have adapted ToBI for other languages and English dialects have focused on reliability agreement between different transcribers as the main evaluation criterion [33,35]. This fact indicates that individual labeling of a single utterance can differ because each transcriber recognizes the prosodic labeling system, 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 agreement between transcribers becomes. However, inter-transcriber agreements in prosodic labeling are often neglected when researchers build and analyze a speech annotated corpus to implement the prosody model.

Related work on linguistics and speech processing has revealed that an accurate realization of breaks depends on the speaker's physical body condition, intention, and utterance habit among native speakers [32,36]. The subject's speaking rate affects the presence and length of prosodic breaks in speech. Even though news announcers are trained and are expected to speak a standard form of the Korean language, the presence and length of prosodic breaks reflect each announcer's style. As shown in Box A2, the realization of prosodic breaks surrounding an adverbial noun such as "oneul" (today) depends on each announcer's speech style. Major breaks were inserted in unexpected places because (1) Announcers focused on the current eo-jeol or the following eo-jeol, (2) they made a mistake, or (3) their speaking rate tended to slow at the end of a sentence. On the other hand, when the speaking rate was faster than usual or when successive breaks appeared, announcers often omitted one.

A single sentence with syntactic ambiguity has several interpretations. In spoken language, prosody prevents garden-path sentences and enables syntactic ambiguity resolution [37,38]. Sentences such as those in Box A3 can be grammatically constructed with multiple syntactic structures. The prosodic phrasing in (a) and (b) can be correct, depending on the sentence's syntactic structure. According to [38], the pattern in Box A3 is quite frequent in Koreans, mainly when the topic is broad. This kind of syntactic ambiguity needs to be resolved using semantic or pragmatic information, as it cannot be resolved using syntactic information only.

As mentioned above, to apply a rule-based approach or a statistical-based approach for predicting prosody phrase boundaries, it is most important to secure a reliable annotation corpus. However, the corpus constructed in previous studies lacks consistency in prosody phrase boundary annotations for various reasons. In this study, we construct a reliable prosody phrase boundary annotation corpus and propose a hybrid method for the prosody phrase boundary prediction method that combines a rule-based approach and a statistical-based approach using the corpus.

The remainder of this study is structured as follows. Section 3 proposes a method for constructing a reliable prosody phrase boundary prediction corpus through coherent annotation. In Section 4, we propose a prosody phrase boundary prediction method that combines a rule-based approach and a statistical-based approach based on the constructed corpus. Section 5 describes the experimental evaluation of the proposed model and com-

pares the results of previous studies with our model. Finally, Section 6 presents a conclusion and some suggestions for further study.

3. Corpus Construction

Based on our analysis of the potential problems in building, collecting, and utilizing a corpus in Section 2, the prosodic breaks cannot be predicted reliably if these problems are not solved. Thus, this section proposes a new method for building a corpus tagged with consistent prosodic breaks among transcribers.

In Section 3.1, a prosodic labeling system that corresponds to the K-ToBI is proposed. In Section 3.2, the selection and preprocessing of the raw corpus are described. To maintain inter-transcriber consistency in tagging prosodic breaks, the overall procedure of training transcribers, individual transcriber tagging, and validation of the reliability of inter-transcriber consistency is illustrated in Section 3.3.

3.1. Selection of Prosodic Labeling System

The definition of prosodic phrases and recognition of prosodic boundaries has been a focus of linguists, phonologists, and speech engineers for a considerable time. As several phonetic phenomena such as consonant assimilation, nasalization, and palatalization occur at the level of the prosodic phrase unit across prosodic words [34], such phonetic phenomena need to be reflected in speech synthesis to implement natural and intelligent TTS systems [16,20].

The K-ToBI was designed for Korea's standard prosodic labeling conventions and was adapted for speech data labeling. In K-ToBI (Korean tone and break indices), four types of break indices, 0, 1, 2, and 3, have been defined [39,40]. However, break indices 0 and 1 cannot be easily discriminated by human ears, nor are they significant to machines. In addition, a prosody labeling system that is simpler, more robust, and easier to use than the complete ToBI system is necessary for the successful training of automatic prosody generation and recognition [41]. Previous studies have divided prosodic breaks into different types according to their needs. Only break and no break were defined in [42–44], whereas three to seven types of prosodic breaks were represented in other systems [20,36,45,46]. Prosodic phrase breaks are usually classified into three types in speech processing: major, minor, and no breaks [20,42].

This study defines six types of prosodic break combined with phrasal boundary tones (because a prosodic break cannot be separated from a boundary tone). In addition, these prosodic breaks are described with syntactic properties that ensure that they are clearly defined, regardless of the subject's speaking rate or style. These six types are defined as follows.

The major break with falling tone: This indicates cases with strong phrasal disjuncture and a strong subjective sense of pause. The positions of major breaks generally correspond to the boundaries of the intonational phrases (marked '///L').

The major break with rising tone: This form is followed by cases with strong phrasal disjuncture but a weak subjective sense of pause length. When there is a short length of a clause, the major break at the boundary between a subordinate clause and the main clause can be weakened. A major break with a rising tone can be inserted when two or more coordinated clauses are connected. A minor break after an adjectival phrase, in contrast, can have a lengthening effect because many words come together and consist of an adjectival phrase (marked '///H').

Minor break with rising tone: This form is followed by 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 a rising tone. Syntactically, when more than one modifier (governed by its subject) appears or more than one argument governed by a predicate appears in the sequence, a minor break is inserted between them. 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').

Minor break with middle tone: This break form is exhibited by 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').

Minor break with falling tone: This form represents cases with minimal phrasal disjuncture and no strong subjective sense of pause. The positions of minor breaks correspond to the boundaries of the accentual phrases with a falling tone. The observations of more raw data revealed that accentual phrase boundaries are sometimes realized in an 'L (low)' tone due to the tonal interaction of adjacent tones and stylistic variations. To date, the detailed characteristics of an AP final L tone and its pragmatic meaning have not been elucidated. 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.

No break: The absence of breaks applies to internal phrase word boundaries. In this case, 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 '#').

In actual data, major breaks with middle tone (or major breaks without tonal change) are observed, although they have no definition or explanation in K-ToBI. Major breaks can replace major breaks with middle tones with rising tone or major breaks with falling tone to implement a prediction model for prosodic breaks because their frequency is low. The syntactic and prosodic analysis of the realization of major breaks with middle tone is complex. The number of major breaks in the data used in the inter-transcriber training and validation experiment was only five, and their ratio was 0.15%. The six types of prosodic breaks are mapped to K-ToBI break indices, allowing further reusability of the corpus labeled by the suggested break types.

In [35], the tonal pattern agreement for each word was approximately 36% for all labelers, and this low-level agreement appears to have been due to the nature of the tonal pattern. Although 14 possible AP tonal patterns exist, these variations are neither meaningful nor phonologically correct. We concluded that the final phrasal tone was sufficient for the recognition of prosodic boundaries. Table 2 shows the relationship between the types of rhyming phrase boundaries proposed in this study and K-Tobi.

Table 2. Mapping between break indices of K-ToBI and the suggested prosodic breaks.

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

3.2. Data Selection and Preprocessing

In this study, TV news scripts were collected as raw corpora. The specifications of the entire raw corpus are listed in Table 3.

Table 3. Information on the source news script data.

Genre	**Source**	**Extraction Method**	**Speaker**
News article	KBS news article scripts	Extraction from web	Female announcer

Although the speech rate of TV news speech is faster than that of generally 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 to the announcer's speech, and their emotional expressions in reading news articles are generally neutralized.

We followed the criteria below for the selection of new sentences.

1. Headline news sentences uttered by one announcer
2. A minimum of five eo-jeols are included in one sentence

First, the text formats of the news scripts extracted from the web were unified. Then, sentences or expressions in news scripts that differed from those in actual sentences in multimedia files were revised according to the natural utterances of the announcer. The revision was performed according to the following criteria.

1. Actual speech of news script read by the announcer was considered as the primary source of prosodic break tagging for labelers.
2. Sentences in the news script were deleted unless the announcer read them in actual speech files.
3. From one to three eo-jeols in news scripts differing from those in speech files were revised according to the actual speech if there was no semantic change.
4. Sentences in the news script considerably differing from those in speech files are deleted.
5. Words or phrases in the news script differing from those in speech files due to spelling/grammar errors are not corrected manually. Spell/grammar errors are corrected automatically by the PNU grammar checker, which shows over 95% accuracy [21].

3.3. Intertranscriber Reliability of Prosodic Phrase Break Labeling

The most reliable method for 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, most 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, is too time-consuming and costly. Most related studies have used a more straightforward method owing to time and cost constraints. If the size of the corpus is small, a professional linguist annotates the entire corpus [47]. If the corpus size is large, more than two transcribers divide the corpus by the number of transcribers, and each transcriber annotates their part [21,48]. Unless the transcribers are trained and the reliability of the inter-transcriber agreement is validated, the consistency of annotation by multiple transcribers cannot be assured. Hence, a method designed to maintain the reliability of the inter-transcriber agreement of prosodic breaks is proposed 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.

First, transcribers read guidelines to familiarize themselves with the prosodic labeling system. Second, to improve the awareness of the length or strength of each prosodic break type in detail, transcribers repeatedly listened to speech files corresponding to several paragraphs in news scripts. In addition, *WaveSurfer*, an open-source program for visualizing and manipulating speech, was utilized for transcribers to examine speech files' pitch contour, waveform, and power plot.

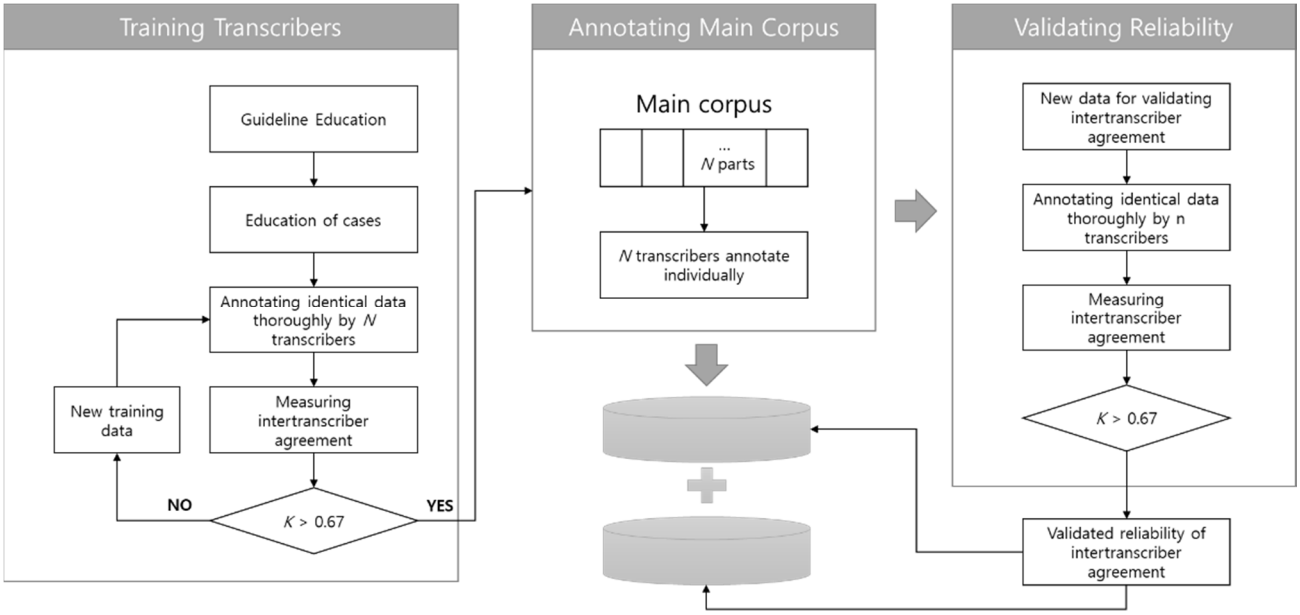


Figure 1. Confirming inter-transcriber reliability of prosodic breaks labeling.

To reduce inconsistency among transcribers, they discussed what they heard and the differences between their perception of the length, strength, or tonal change of a single prosodic break. After mastering the guidelines and training with the above mentioned examples, specific reasons for inconsistency among transcribers were analyzed, and their solutions were suggested 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-jeols from the end of a sentence. Some transcribers recognized these as speaker errors and corrected them in their annotations. On the other hand, others annotated prosodic breaks according to what they heard, regardless of the errors. Due to these differing policies on annotation, the resultant annotation of prosodic breaks among transcribers was not consistent. Inconsistencies derived from these speaker errors should be deleted.
2. If the speech rate of the announcer was too fast for some transcribers to perceive audible breaks between two eo-jeols, they omitted the minor break, whereas others recorded a minor break in the same place. In this case, transcribers need to pay attention to whether the final tone of the target eo-jeol rises or falls. To reduce inconsistency derived from missing breaks, transcribers repeatedly practiced while listening to similar patterns.
3. If only one annotator selected a different type of prosodic break than the others for the answer of the same place, they were required to change their annotating prosodic breaks.
4. Several previous studies have revealed prosodic variability even for news speech data [31,37]. The announcer showed variability in the location, strength, or length, and tonal change in our news data. For example, the announcer occasionally inserted a minor break between two eo-jeols consisting of a time expression.

Five transcribers annotated the same data with prosodic breaks simultaneously and then compared the results of their annotations and discussed and repeatedly corrected the various errors until reliable agreement among them was reached. The inter-transcriber agreement in annotating six-level prosodic breaks, including tonal changes, is shown in Table 4.

Table 4. Inter-transcriber agreement in tagging prosodic break with tone.

Agreement	Cumulative Rate of Agreement			
	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%

The cumulative agreement rate of more than half of the transcribers ($n + 1/2$) was measured using approximate figures. Precisely, the rate of the inter-transcriber agreement was 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 was relatively low, although the first experiment was performed after the transcribers had familiarized themselves with the guidelines and studied many examples. The inter-transcriber agreement in annotating data with six-level prosodic breaks increased continuously with repeated training and experiments. This indicates that educating transcribers with guidelines and examples is not sufficient, and training of transcribers is required before the annotation of the main corpus with specified tagging classes by multiple transcribers.

The inter-transcriber agreement in annotating three-level prosodic breaks excluding tonal changes showed higher agreement than that in annotating six-level prosodic breaks, as shown in Table 5.

Table 5. Inter-transcriber agreement in tagging prosodic break alone.

Agreement	Cumulative Rate of Agreement			
	1st	2nd	3rd	4th
Five (all) agreed	51.42%	57.17%	60.89%	62.73%
At least four agreed	69.67%	80.14%	83.71%	83.69%
At least three agreed	91.94%	95.40%	96.54%	95.36%

The rate at which at least three of the transcribers agreed and at which at least four agreed did not increase further in the fifth experiment, whereas the rate at which all of them agreed increased slightly. The annotation accuracy of each transcriber was estimated To review how accurately each transcriber annotated the corpus, the prosodic break type for which at least three of them agreed was considered as the correct answer. The annotation result of each transcriber was compared to the answer, and the accuracy was estimated by counting the number of annotations that matched the answers. Table 6 shows the estimated annotation accuracy of the five transcribers from the first to the fourth experiment.

Table 6. Estimated annotation accuracy of an individual transcriber.

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

Although there are individual variations, the estimated accuracy of the transcribers increased steadily. After the four experiments, the cumulative agreement rate 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 inter-transcriber agreement is required to determine whether the training is sufficient. The reliable measurement of the inter-transcriber agreement was initially studied earlier [33], as the goal of the original ToBI system designers was to design a system with the following features.

1. Reliability: agreement between different transcribers must be at least 80%.
2. Coverage: sufficiently comprehensive coverage to capture the most critical prosodic phenomena in spontaneous speech is required.
3. Learnability: to be used in multi-site data collections, training time must be relatively short.
4. The capability of being related to current approaches to speech recognition, parser outputs, and formal representations of semantics and pragmatics is required.

The designers and developers of adaptations of ToBI for other languages and dialects such as K-ToBI, G-ToBI, J-ToBI, and GlaToBI have proven the usability of their system based on the above mentioned criteria. They have also studied the measurement of inter-transcriber agreement [33,35,49].

The most commonly used methods to assess agreement among transcribers are pairwise analysis and kappa statistics. The pairwise analysis evaluates the original ToBI system and the version developed for German [49,50]. This method compares the labels of each transcriber with the labels of every other transcriber for that particular aspect of the utterance [33]. The basic unit for measuring agreement is the transcriber-pair-word or the set of two labels assigned to one word by a pair of transcribers. Inter-transcriber consistency is the percentage of transcriber-pair-words exhibiting agreement on a particular element in the transcription.

$$\frac{a(a-1)}{n(n-1)}. \quad (\text{number of transcribers: n, number of transcribers that agreed: a}) \tag{1}$$

Carletta (1996) adapted the Kappa coefficient of agreement (K) suggested in [51] to assess the reliability of inter-transcriber agreement in prosodic annotation [52]. K is "the ratio of the proportion of times the raters (transcribers) agree (corrected for chance) to the maximum proportion of times that the raters (transcribers) could agree (corrected for chance)" as given in the following equation.

$$K = \frac{P(A) - P(E)}{1 - P(E)}, \tag{2}$$

where $P(A)$ is the proportion of times that the transcribers agree (i.e., the percentage agreement from the pairwise agreement above) and $P(E)$ is the proportion of times that the same number of transcribers agree by chance.

According to Carletta (1996), the rate of agreement of transcribers expected by chance depends on the number and relative proportions of the categories used by the transcribers; if there are only two available categories, both of which have an equal chance of occurring, then two transcribers using these categories agree 50% of the time; if the number of categories is increased to four, the chance of agreement is 25%. Thus, she suggests that Kappa statistics, which consider both the number and proportion of categories and the chance of agreement, are more helpful in judging reliability. In addition, she states that "the interpretation of the scale of agreement is possible". Values of $K > 0.8$ indicate good reliability, while values of $0.67 < K < 0.8$ indicate possible reliability.

The main corpus comprising 29,686 eo-jeols was divided into five parts. Each partition is assigned to five trained transcribers, and annotation was performed independently. *WaveSurfer*, which was used in the training phase, is also used in the annotation phase to display and annotate speech. Transcribers may openly discuss their annotations, even

though they annotated different parts of the main corpus. Because each transcriber annotated a different part of the main corpus, the reliability of the inter-transcriber agreement could not be measured directly. We assumed that the inter-transcriber agreement did not change significantly before and after the annotation of the main corpus.

Hence, another dataset, including 1149 eo-jeols (46 sentences), with a size 1.5× that of the dataset used in the 4th experiment, was collected and used instead to validate the reliability of the agreement. Immediately after annotation of the main corpus, the final experiment was 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 were used to measure inter-transcriber agreement on the validation data set. The reliability of the validation experiment is presented in Table 7. The pairwise agreement and K found in the validation experiment after annotation of the main corpus were 0.79 and 0.76, respectively.

Table 7. Reliability of inter-transcriber agreement on the validation set.

Measurement	Reliability of Inter-Transcriber Agreement
The cumulative rate of at least three agreed	0.9295
Pairwise analysis	0.7916
Kappa statistics	0.7569

Both agreement figures in Table 8 are greater than those found in the prior experiments, which were repeated four times during the training phase. Based on this result, the main corpus's annotation is also considered part of the training of the transcribers. The estimated pairwise agreement of annotation of the main corpus was between 0.7477 and 0.7916, and the value of K was between 0.7057 and 0.7569. Considering the estimated K, the annotation of the main corpus had reliable consistency among multiple transcribers. As a result, we obtained a corpus with a consistent annotation of prosodic breaks. The main corpus was divided into an analysis corpus and two sets of evaluation corpora, as shown in Table 8.

Table 8. Constructed corpus with consistent prosodic labeling.

Purpose	Characteristics	No. of Eo-Jeols	No. of Sentences
Analysis corpus	Data set from validation experiment	1149	46
	80% of main corpus	24,185	1092
Evaluation corpus 1	10% of main corpus	2613	109
Evaluation corpus 2	10% of main corpus	2865	118
Total		30,812	1365

The corpus annotated by all transcribers in the validation phase was also added to the analysis corpus. In this section, methods for training transcribers, annotating a corpus using multiple transcribers, and validating the reliability of inter-transcriber agreement have been suggested. Following the overall procedure, a corpus including 30,812 eo-jeols was constructed, and the reliability of its inter-transcriber agreement was validated, although there were time and expense limitations.

4. Prediction of Prosodic Breaks Using Deep Learning and Rules

Section 3 has described the proposed method to construct a reliable prosody boundary annotation corpus to develop prosody boundary prediction technology for natural speech synthesis. This section describes our proposed method to use a corpus to analyze patterns

on prosody boundaries, create rules, and use them for training deep learning-based prosody prediction models.

4.1. Prediction of Prosodic Breaks Using Rules

Related work in linguistics and speech processing has reported that the realization of breaks strongly depends on the speaker's physical body condition, intention, utterance habit, and speed, even among native speakers. Despite the variable characteristics of prosodic breaks in everyday speech among native speakers, communication of meaning in conversation is rarely inaccurate. Although some variable factors exist, native speakers share common concepts and rules for generating prosodies. Past reviews have determined that prosodic and syntactic structures are strongly related. The boundaries of syntactic phrases can provide essential clues for predicting appropriate prosodic breaks because prosodic chunks are semantic units in an utterance [36]. For example, prosodic breaks are realized in the middle of two different syntactic phrases, as illustrated in Figure 2.

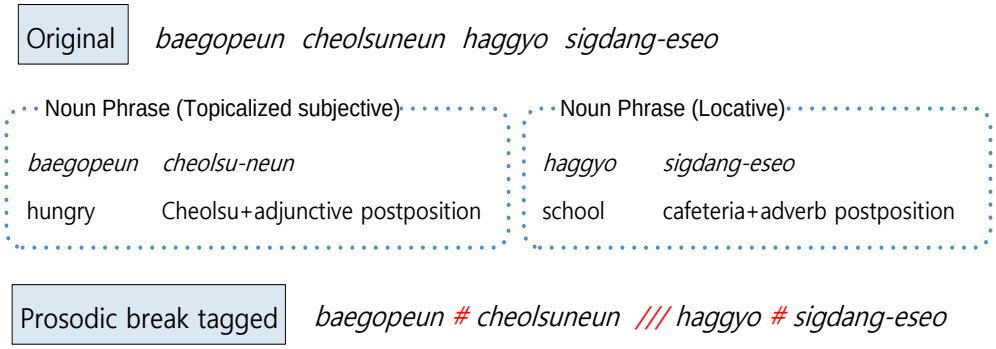


Figure 2. Prosodic break realization in boundaries of syntactic phrases.

In this sense, the best way to find the correct placement of prosodic breaks in a sentence is to utilize a parser that can detect syntactic boundaries accurately. However, implementing a high-performance parser presents a new challenge in natural language processing, and there are no full parsers available for Korean. However, constituents in a previous prosodic phrase do not affect the realization of a break after the current word. For example, in Figure 2, "baegopeun" modifying "Cheolsu" does not provide any information as to whether or not a break occurs after "haggyo".

Thus, in this study, the syntactic boundaries and syntactic relations between constituents were obtained by partial parsing. A morphological analyzer and a previously developed POS-tagger [53] were used in this study to obtain syntactic information from text. The POS-tagger exhibited an accuracy of 96.8% by adopting a stochastic tagging method and generating accurate POS tag sequences from Korean sentence inputs. However, a POS sequence alone cannot provide sufficient clues for detecting all syntactic phrase boundaries. To address this problem, we collapsed the initial 43-tag set used in the POS tagger. We expanded it to include sub-categorized common nouns, predicates, adverbs, and conjunctive endings depending on syntactic function.

Sub-categorization of nouns by their syntactic function: Some nouns modify the following nominal words or phrases. For example, nouns describing a certain status such as "*jaebul* (non-French residents)," "*jaeil* (non-Japanese residents)," "*wanjeon* (perfect)," and "*sagyojeog* (sociable)" cannot be used alone. Instead, they should precede and modify other nouns to form compound nouns. These are adnominal nouns (AdnN). Nouns which denote time such as "*hyeonjae* (present)," "*oneul* (today)," and "*jinanhae* (last year)," or indicate degrees such as "*daso* (a little)," "*choedaehan* (best, at full bore)," "*choesohan* (at the very least)," "*jajin* (voluntary),' and "*jeoggeug* (positively, actively)" perform an adverbial function and are thus known as adverbial nouns (AdvN). Several nouns known as predicative nouns (PrdN) perform a verbal function or predicative adjectival function. Examples

include "*chulje* (making questions)," "*seonbal* (selection)," and "*gae-ib* (intervention)." The remaining noun were nominal nouns (N).

Sub-Categorization of Predicates: Korean verbs and predicative adjectives are sub-categorized into different syntactic categories. They express their semantic arguments using different syntactic means, called a sub-categorization frame [54]. <Predicate, case frame> pairs in electronic dictionaries created as a subdivision of the 21st Sejong Project have been used to subcategorize Korean predicates in the past.

Sub-Categorization of Adverbs: Adverbs are classified as general adverbs that modify the composition of a sentence or conjunctive adverbs that connect the sentence components. A general adverb can again be sub-categorized according to the component that it modifies. These sub-categories are adverbs that modify nouns, adverbs that modify adjectives or adverbs, and adverbs that modify verbs and adverbs that modify clauses. A total of 3112 adverbs were classified as one of these sub-categories and were included in the POS sub-categorization dictionaries. The prosodic phrase break can be determined between an adverb and the following word using the sub-category information of an adverb. For example, if an adjective follows an adverb that modifies an adjective, there would be no break between the two words.

Sub-categorization of conjunctive endings: Conjunctive endings are used to link sentences, clauses, or predicates. In the Korean language, predicates are positioned at the end of clauses, and conjunctive endings are mixed, making it difficult for a system to detect clausal boundaries. Conjunctive endings were classified into two groups depending on their syntactic function to reduce the complexity of finding clausal boundaries. One group encompasses clausal conjunctive endings, and the other group consists of the remaining conjunctive endings that connect predicates and clauses. Immediately next to a clausal conjunctive ending, a syntactic clausal boundary always appears. A conjunctive ending, '*-ge*,' is combined with predicative adjectives and converts the syntactic characteristic of predicate adjectives into adverbial modifying verbs in Korean.

In the Korean language, word order is more complementary than English, and syntactic constituents such as a subject or an object and case markers are frequently omitted. Thus, dependency grammar is used for analyzing syntactic relations herein, and the dependency relations between the constituents are given in Table 9.

Table 9. Dependency relations between constituents.

Governor	Specific Dependency Relation	Dependent
Nominal noun	modification	Pronoun, adjectival noun, general/numeral modifier, modifier ending, modifier proposition
	connection	The connective proposition, connective adverb
Pred	argument	The subjective proposition, objective proposition, noun, nominalization ending
	modification	Adverb modifying verbs, adverbial clausal ending
	connection	Conjunctive endings,
Modifier	modification	Adverb modifying adjectival/adverbial
Postposition	case assignment	Noun, pronoun, number, nominalization ending

Prosodic breaks determined by the co-relation between syntactic structure and prosodic structure are formalized as rules for implementing the prediction system of prosodic breaks. The conditions of the rules consist of a combination of the subclassified POS of a target word or its contextual words as shown in Algorithm 1.

Algorithm 1 Prediction of prosody phrase boundary using syntactic information.

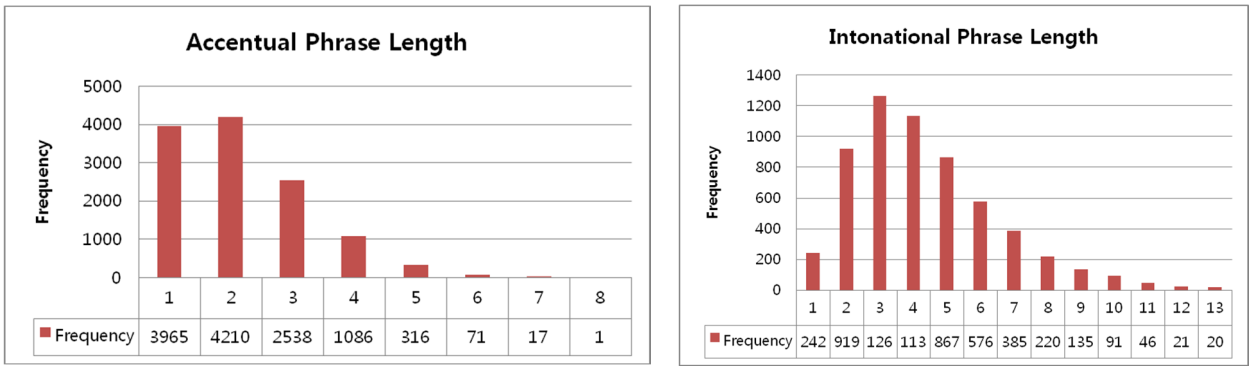
```
Input: Sentence
Output: Sentence with prosody phrase boundary
begin
  for each eo-jeol in the sentence do
    if end POS of current eo-jeol == "interjection" then
      Phrase_Break[current_position] = "///H"
    else if start POS of next eo-jeol == "adjectival noun" then
      Phrase_Break[current_position] = "//H";
    ...
    else then
    Phrase_Break[current_position] = "#";
end
  end
end
```

4.2. Feature Extraction for Prediction of Prosodic Breaks

The learning features mainly used in previous studies include (1) part-of-speech information of an eo-jeol, (2) the length of an eo-jeol, and (3) the distance from a specific position to the current eo-jeol. In previous studies, the distance feature is the distance from the beginning or end of a sentence to the current eo-jeol, its normalized value, the distance from the previous punctuation included in the current sentence to the current eo-jeol, and the distance from the nearest preceding dependent/dominant to the current eo-jeol. This was extracted in various ways, such as by distance. However, the longer the sentence, the more critical is the distance from the prosody boundary that occurred before the current eo-jeol rather than the distance from the beginning/end of the sentence to the current eo-jeol. In addition, the usage of punctuation marks is very ambiguous, and the pronunciation and the occurrence of pauses vary depending on the usage, so the information obtained from these qualities is often meaningless. Another disadvantage is that the distance from the dominance/dependency depends heavily on the parser's performance. In this study, we propose an extraction method to use distance information and collocation information as learning qualities along with the part-of-speech information of the word used in the rule-based model.

Related work in linguistics and speech processing has reported that the realization of breaks strongly depends on the speaker's physical body condition, intention, utterance habit, and speed, even among native speakers [36]. Because of the variable characteristics of prosodic breaks, major versus minor break prediction differences are not regarded as fundamental errors in the sense that the hand-labeled prosodic boundaries are mismatched for the exact text [20]. Therefore, it is necessary to predict irregular prosody boundaries through information other than linguistic information, such as part-of-speech information or syntax information.

Some studies have reported that phrases occur at somewhat regular intervals and that the likelihood of a break occurring increases with distance from the last break [16,18,45]. English intonational phrases include three to six words. According to [43], the Korean AP includes five or fewer syllables at a standard speech rate and can include up to seven syllables at faster rates. The distribution of accentual phrase lengths in eo-jeols and the intonational phrase length in eo-jeols are shown in Figure 3, respectively. In our news data, most phrases were between one and three eo-jeols long, whereas most intonational phrases were between two to five eo-jeols long.



(**a**) (**b**)

Figure 3. Distribution of prosody phrase length in eo-jeols. (**a**) Distribution of accentual phrase length in eo-jeol. (**b**) Distribution of intonational phrase lengths in eo-jeols.

Previous work has attempted to use distance information, but existing data-driven learning models cannot ascertain where the previous break occurred. Therefore, it is impossible to compute the distance from the previous brake reliably. Some researchers have simply computed the distance in words from the beginning and end of each sentence and the distance in words from the previous internal punctuation to the current word [18]. This results in definitive placements, but not in a highly related position to the current break. In addition, punctuation is ambiguous in texts, as described in [19,55,56]. Ref. [19] uses the distance from the current eo-jeols to its governor in syllables and eo-jeols. Because no reliable and high-performance syntactic parser has been developed, it is difficult to obtain reliable results from syntactic parsing in many languages, including Korean. The author of [48] claims that these approaches are not recommended because a single error cannot be recovered from, which may cause errors in all subsequent decisions. Instead, they proposed an n-gram model to estimate the probability of a break at any of the previous junctures. They also examined all the possibilities and thus found the most likely sequence of the juncture type for the input POS sequence.

Successive non-breaks were more excellent than after applying rules established in Section 4 occurred, and restrictions on successive non-breaks were triggered using POS bi-gram pattern rules. Each bi-gram comprised the POS of the last morpheme of the current eo-jeol and POS of the first morpheme of the next *eo-jeol* bi-gram. The likelihood ratio measured the relatedness of the bi-gram patterns and the prosodic break type.

Words do not appear randomly in sentences but rather tend to appear together with other words. This phenomenon in which words frequently appear together is called collocation. Here, "frequently together" does not simply mean more or less in absolute frequency in the corpus, but whether it is more than expected or not is the key to forming a collocation relationship. Whether these two words form a collocation can be used as a quality for predicting the prosody boundary. In this paper, hypothesis testing determined whether or not collocation is formed between the first morpheme of the previous word and the last morpheme of the next word based on the distance between words. This is because the type and position of the boundary are mostly determined by the first and last morphemes of the word. To determine whether two morphemes form a collocation using the test of independence, the hypothesis to be tested should be established as follows. Two assumptions are made regarding the relatedness of the morpheme bi-gram(C_{i-1} C_i) and each prosodic break (PB_k).

Hypothesis 1 (H1). $P(C_{i-1}C_i|PB_k) = p = P(C_{i-1}C_i|\neg PB_k)$.

Hypothesis 2 (H2). $P(C_{i-1}C_i|PB_k) = p_1 \neq p_2 = P(C_{i-1}C_i|\neg PB_k)$.

Hypothesis 1 indicates that the occurrence of C_{i-1} C_i is independent of the occurrence of a prosodic break, PB_k. Hypothesis 2 is a formalization of dependence (the occurrence of PB_k is dependent on the occurrence of C_{i-1} C_i). Here, cnt$_1$ is the count of C_{i-1} $C_{i,}$ and cnt_2 is the count of PB_k occurrence. cnt$_{12}$ is the co-occurrence of the bi-gram pattern and prosodic break. P can be calculated using the maximum likelihood estimation as follows.

$$p = \frac{cnt_2}{N},\ p_1 = \frac{cnt_{12}}{cnt_1},\ p_2 = \frac{cnt_2 - cnt_{12}}{N - c_1}, \tag{3}$$

where N is the number of eo-jeols in the corpus.

One obtains the likelihoods $L(H_1)$ and $L(H_2)$, and the likelihood ratio, λ, is given in Equation (4).

$$\lambda = \frac{max_{\omega \in \Omega_0} H(\omega; k)}{max_{\omega \in \Omega} H(\omega; k)}, \tag{4}$$

where ω is one point in the parameter space, Ω and Ω_0 is the sample space in the parameter space according to the hypothesis. Assuming a binomial distribution, Equation (4) can be interpreted as Equation (4). The log of the likelihood ratio λ is expressed as follows.

$$\log \lambda = \log \frac{L(H_1)}{L(H_2)} = \log \frac{b(cnt_{12}, cnt_1, p) b(cnt_2 - cnt_{12}, N - c_1, p)}{b(cnt_{12}, cnt_1, p_1) b(cnt_2 - cnt_{12}, N - c_1, p_2)} \tag{5}$$

$$L(H_1) = b(cnt_{12}, cnt_1, p) b(cnt_2 - cnt_{12}, N - c_1, p), \tag{6}$$

$$L(H_2) = b(cnt_{12}, cnt_1, p_1) b(cnt_2 - cnt_{12}, N - c_1, p_2). \tag{7}$$

By applying Equations (6) and (7), Equation (5) is then converted into Equations (8) and (9):

$$\begin{aligned} \log \lambda = &\log L(cnt_{12}, cnt_1, p) + \log L(cnt_2 - cnt_{12}, N - c_1, p) \\ &- \log L(cnt_{12}, cnt_1, p_1) - \log L(cnt_2 - cnt_{12}, N - c_1, p_2), \end{aligned} \tag{8}$$

$$L(k, n, x) = x^k (1 - x)^{n-k}. \tag{9}$$

Because the quantity $-2 \log \lambda$ is asymptotically x^2 distributed, we can use the value to test the null hypothesis H1 against the alternative hypothesis H2. The critical value for a single degree of freedom is 7.88 in a x^2 distribution. Now, we can look up the value of 1238.28 for the bi-gram pattern, 'topical postposition-adjective' and the prosodic break '//H' and reject H1 for this bigram on a confidence level α = 0.005.

4.3. Bidirectional LSTM-CRF-Based Prosody Boundary Detection

In this study, rule-based prosody boundary prediction results were applied to a deep learning technique, which has recently shown excellent performance in natural language processing. In particular, a bidirectional LSTM CRF classifier [29,30,57], which has shown excellent performance in the sequence labeling problem, was used. This section introduces the structure of the deep learning model and the data and features used in the model for an overall description of the proposed model.

Figure 4 shows the overall structure of the proposed bidirectional LSTM-CRF-based prediction model in word units for prosody boundary prediction. The first layer processes the distributed representation of words and maps input words into word vectors for processing in subsequent layers. Then, a bidirectional LSTM-CRF-based layer predicts the prosody type (major, minor, and no break) corresponding to each input word. In the example in Figure 4, "baegopeun # chcheolsuneun///haggyo # sigdang-eseo///" a major break occurs only in "chcheolsuneun" and "haggyo", and the reading is interrupted.

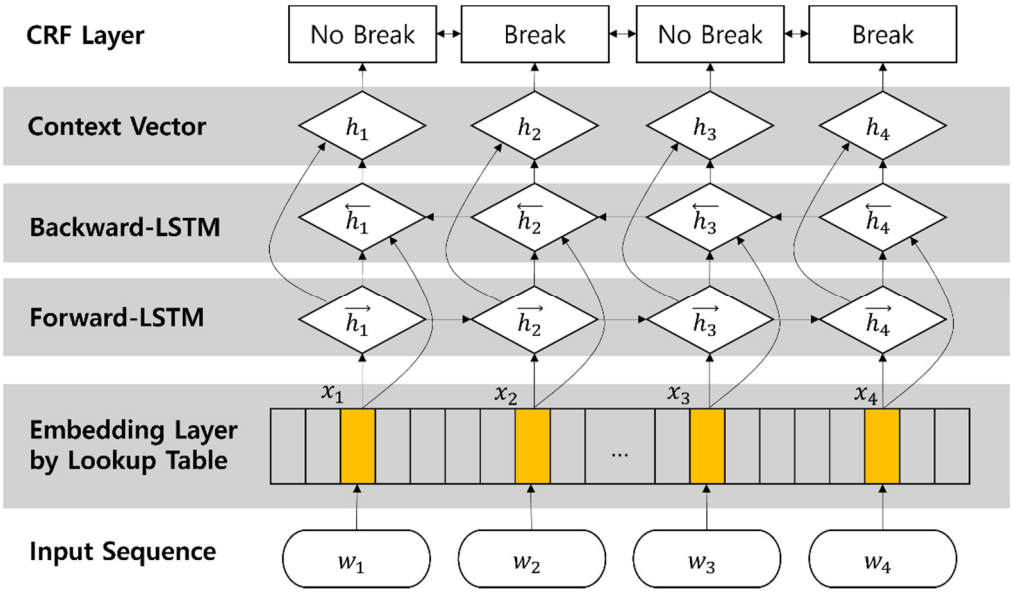


Figure 4. Structure of BLSTM-CRF modeling.

In a previous study, to efficiently utilize the part-of-speech information, a more subdivided part-of-speech set was used than the general part-of-speech set used in morpheme analysis. In addition, when using the part-of-speech information of the word, it was found that using only the part-of-speech information of the first morpheme and the last morpheme, which are more involved in predicting the prosody boundary, was effective in predicting the prosody boundary. To reflect this in the bidirectional LSTM-CRF, 406-dimensional word embeddings were constructed, as shown in Figure 5.

	200D	200D	3D			2D	
Input	Word Embeddings		Prosody phrase boundary predicted by the rule-based model			Features	
	first morpheme	last morpheme	Major Break	Minor Break	No Break	Phrase Length	Collocation
baegopeun	[...]	[...]	0	0	1	1	0
chcheolsuneun	[...]	[...]	0	1	0	2	0
haggyo	[...]	[...]	0	0	1	1	0
sigdang-eseo	[...]	[...]	0	1	0	2	0

Figure 5. Structure of eo-jeol embeddings.

Each word has a 200-dimensional size consisting of the vector of the first and last morpheme, the boundary type predicted by the rule-based model, the distance from the boundary predicted by the rule-based model to the current word, and whether or not collocations are formed. This word embedding method enables the combination of a rule-based model and a statistics-based model.

5. Experiments and Evaluation

5.1. Experimental Data

Previous studies' prosody phrase boundary types are usually level 2 (break/non-break) and level 3 (major break/minor break/non-break). The more the subdivided prosody phrase boundary type is used, the more complex the system determines which class of the subdivided types to classify. The disadvantage is that the tagging consistency and accuracy between the annotators in the prosody phrase boundary tagging step decreases. On the other hand, when the system correctly predicts, it can provide usefully divided information, and it can improve the clarity and naturalness of the synthesized

sound generated by the speech synthesis system. We separated the test data from the analysis data and used it for all the experiments described below. The average length of a sentence is 24.27 eo-jeols, which is rather long. Table 10 presents the distribution of the four types of prosodic breaks in the experimental data.

Table 10. Distribution of the five types of prosodic break in the experimental data.

	Train	**Test**
///H	2181(10.2%)	241(9.8%)
///L	2514(11.8%)	296(12.1%)
//H	4964(23.2%)	599(24.5%)
//M	15(0.1%)	1(0.0%)
//L	55(0.3%)	5(0.2%)
#	11,653(54.5%)	1305(53.3%)
Total	21,382	2447

5.2. Evaluation Measure

Although no single method to measure the performance of a prosodic break prediction algorithm has been established, a variety of approaches have been proposed in previous studies [48]. To explore the multifaceted evaluation of the proposed system, various performance criteria suggested in previous studies were adopted in this study. Two methodologies for evaluating a prosodic break prediction system were developed. First, the performance was measured by matching the predicted value for every word boundary with the corresponding values tagged by trained annotators. Precision, recall, and f-measure in [10,35] estimated the overall system performance. Precision is defined as the number of correctly identified instances of a class (*tp*) divided by the number of correctly identified instances (*tp*) and the number of wrongly selected cases (*fp*) for that class. The recall was estimated as the number of correctly identified instances of a class (*tp*) divided by the number of correctly identified instances plus the number of cases the system failed to classify (*fn*). The f-measure is the harmonic mean of the precision and recall, calculated as

$$P = \frac{tp}{tp + fp}, \tag{10}$$

$$R = \frac{tp}{tp + fn}, \tag{11}$$

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}, \tag{12}$$

where α is a factor that determines the weighting of the precision and recall. A value of $\alpha = 0.5$ is often used for equal weighting of precision and recall.

Performance was assessed with reference to the total number of word boundaries in the test set (N), and the total number of word boundaries that were assigned as breaks in the test set (B). A deletion error (D) occurs when a break is marked in the reference sentence but not in the test sentence. An insertion error (I) occurs when a break is marked in the test sentence, but not in the reference. A substitution error (S) occurs when a break occurs in the right place but is of the wrong type. This type of error is only relevant when more than one type of break is considered. These performance measures were calculated as follows.

$$\text{Break\_correct} = \frac{\text{B} - \text{D} - \text{S}}{\text{B}} \times 100\%, \tag{13}$$

$$\text{Non\_Break\_correct} = \frac{\text{N} - \text{I} - \text{S}}{\text{N}} \times 100\%, \tag{14}$$

$$\text{Non\_Break\_correct} = \frac{N - I - S}{N} \times 100\%, \tag{15}$$

$$\text{alse insertions w.r.t juntures} = \frac{I}{N} \times 100\%, \tag{16}$$

$$\text{False insertions w.r.t breaks} = \frac{I}{B} \times 100. \tag{17}$$

According to [48], the difference between the break-correct and juncture-correct depends on whether non-breaks are included in the calculation. Note that while the break-correct score only gives credit to breaks correctly predicted, the junctures-correct score accounts for both correctly predicted breaks and non-breaks and is therefore sensitive to the ratio between breaks and non-breaks in a text. In data from [48], the number of non-breaks outnumbered the number of breaks by a ratio of approximately 4:1; hence, an algorithm that marked everything as a non-break would score approximately 80% as juncture-correct, but 0% break-correct.

5.3. Evaluation Results

For performance comparison according to the level of boundary type, Break_correct, Juncture_correct, and Adjusted_score values were used to evaluate the performance of the proposed system. Table 11 shows the overall performance of the proposed system.

Table 11. Overall performance of suggested system.

Level	Performance Measurement			
	Accuracy	**Break_Correct**	**Juncture_Correct**	**Adjusted_Score**
6 level	75.96%	58.31%	75.97%	0.48
3 level	78.57%	63.99%	78.57%	0.53
2 level	87.33%	83.12%	87.33%	0.72

As expected, it can be seen that the more complex the prosody sphere boundary type, the lower the prediction performance. In particular, at level 6, it can be seen that the learning was not performed properly because the data of//M and//L were not enough. Therefore, it is more efficient to use a three-level rhyme boundary even for practical purposes.

We compared the two models proposed in previous studies to evaluate the performance of the model proposed in this paper. The first is a model that combines GRU-based bidirectional RNN and multi-head attention (Bi-GRU-Attention) [23,26,27], and the second is a model that applies fine-tuning based on BERT.

Figure 6 shows the overall structure of a GRU-based bidirectional recurrent neural network and multi-head attention-based model for prosody phrase boundary prediction. It is a model that predicts whether a prosody boundary occurs between each sentence word when a sentence is an input. This model consists of an input layer using word2vec, a GRU-based bidirectional recurrent neural network, self-attention-based multi-head attention, a highway network, a fully connected layer, and an output layer. A 200-dimensional morpheme vector was learned skip-gram model using the learning data separated by 9:1 from the 'UCorpus-HG corpus,' a morphological, semantic analysis corpus of about 18.87 million words. Multi-head attention is a form in which dot product attention is superimposed, and a scale with scaling is added in the middle through a linear layer by dividing feature values for query, key, and value.

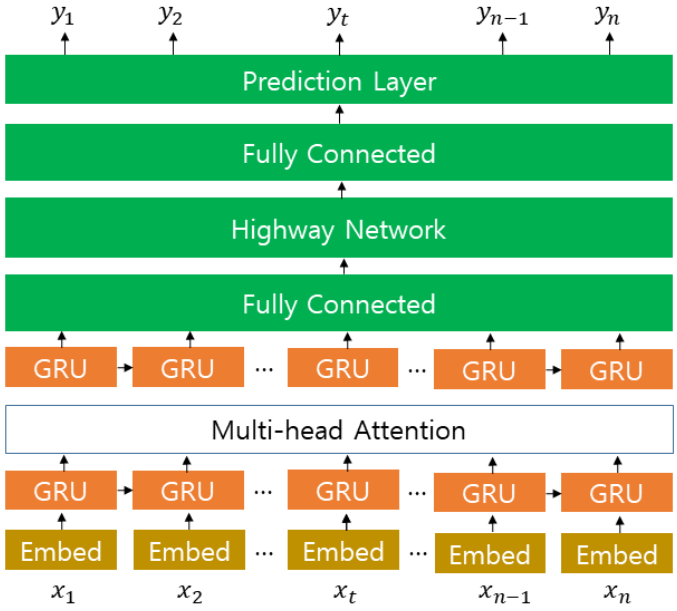


Figure 6. Structure of Bi-GRU-Attention-based rhyme sphere boundary prediction model.

BERT is a model trained using the transformer used in the GPT model. Unlike GPT, which proceeds in one direction, learning proceeds in both directions. When learning is conducted in both directions, an indirect reference to oneself occurs, and BERT uses the Masked LM technique to solve this problem. In this study, only the shape of the output layer is modified and applied to the prosody sphere boundary prediction model through fine-tuning that shares the critical parameters of BERT. However, we converted the input into syllable units in the input layer, and the boundary between syllables within a word was set as a prosody non-boundary.

As can be seen from the results in Table 12, the BLSTM-CRF model proposed in this study exhibited higher accuracy in predicting the boundary of Korean prosody than the CRF model learned with various learning qualities. However, these approaches based on learning have several disadvantages, including (1) if there is no established or shared tagged corpus appropriate for a research goal, a tagged corpus must first be constructed, and (2) the speech annotated corpus is expected to be sufficiently large and reliable to obtain accurate statistics and learning results, because machines cannot discriminate meaningful data from noise. There are few large-scale corpora currently available to the public, and they have various inconsistency problems that can result in problematic learning, as we have seen in Section 2. In addition, POS sequences alone are insufficient for detecting syntactic structure, which is strongly related to prosodic structure.

Table 12. Overall performance comparison with previous work.

Model	Model		
	BERT	**Bi-GRU-Attention**	**Proposed Model**
Major break	0.9832	0.9601	0.9268
Minor break	0.6191	0.5847	0.6928
No break	0.7608	0.7695	0.9597
Macro-Avg-Precision	0.7877	0.7714	0.8663
Micro-Avg-Precision	0.9522	0.8976	0.9437

Our system predicted prosodic phrase breaks with relatively higher performance due to the analysis of the syntactic structure. The performance of our system was particularly excellent in the prediction of major breaks and non-breaks. However, our system did not

perform significantly better in the prediction of minor breaks. This occurred due to the variable nature of the minor break, which can be either shortened or lengthened according to various semantic and pragmatic factors. To predict variable prosodic phrase breaks, other elements such as word sense, collocations, focus, and the intention of the speaker need to be considered.

Finally, we evaluated whether the proposed model contributes to improving the quality of speech synthesis service in the medical field. In speech synthesis, MOS (mean opinion score), the most common method for measuring speech quality, was conducted. Three participants ('30 s, '40 s, '50 s) were told the voice to which the rhyming phrase boundary prediction was applied and the voice without the rhyming phrase boundary prediction, and they were asked to give a score of 1 to 5 for pronunciation, intonation, speed, and break. The speech synthesis system used Microsoft's Azure. Because Azure can receive SSML files as input, rhyming phrase boundary prediction can be applied to the input sentence. By predicting the three-level prosody phrase boundary, we gave a break of 2 s for a major break and 1 s for a minor break. The voices to be heard by the participants read ten sentences included in the instructions included in Company A's headache medicine, an over-the-counter medicine.

As shown from Table 13, the total score was 3.41 for the primary voice and 3.32 for the voice, including the prosody boundary prediction result, giving a higher score to the primary voice. However, speed or interrupted reading gave higher scores for speech, including prosody boundary prediction results. Interestingly, although the two voices differed in the brakes, the participants gave different evaluations in pronunciation, intonation, and speed. In the future, we plan to increase the reliability of the evaluation by conducting evaluations with more diverse participants.

Table 13. Comparison of MOS evaluation results.

	Speech Synthesis Result of Text Including Prosody Phrase Prediction result	Speech Synthesis Result of Plain Text
pronunciation	3.36	3.58
intonation	3.19	3.40
speed	3.44	3.39
break	3.30	3.27
Total	3.32	3.41

6. Conclusions and Future Work

In this study, various components of a sentence were sub-categorized, and syntactic information was utilized to predict the location of natural prosodic phrase breaks. The correlation between the analyzed syntax structure and prosody structure established rules for predicting prosodic phrase breaks. As for the overall accuracy in predicting total prosodic phrase breaks, our system elicited a Break_Correct result of 63.99% and a Juncture_Correct result of 78.57% in predicting three levels of prosodic break. In addition, this study proposed an effective method for combining a rule-based model and a deep learning-based model. When there is not enough training data, it has been proven through experiments that combining a rule-based model and a deep learning-based model is more effective than using a deep learning-based model alone. However, additional factors in determining optional prosodic breaks must be considered regardless of the syntactic structure. The evaluation of optional prosodic breaks at the sentence level should also be considered.

The following contributions of the study would be helpful in related future work.

1. Potential problems in the construction of the speech annotation corpus have been identified, and a solution for each type of problem has been suggested. The overall

procedure of training transcribers has also been described, and the results of inter-transcriber agreement training have been presented.
2. Rules using dependency relations between syntactic constituents have been established. These rules can predict mandatory prosodic breaks and correct errors in the annotated data, which data-driven models cannot recognize.
3. We combined the probabilities and rules for reliable and flexible prediction of optional breaks. Optional breaks can be identified when the speaking rate is set at a slow speed, whereas these breaks are shortened or disappear when fast utterances are required from the systems. This characteristic of the suggested model is beneficial for implementing flexible TTS systems that can generate natural prosody of sentences.
4. The implemented system can straightforwardly process input sentences. Distances from the last mandatory position of the prosodic break predicted by rules are stored and computed dynamically to predict the break type of the current eo-jeol. In this way, the proposed system can be adapted to real-time TTS systems.

One limitation of this study is that no learning data at scale in the medical or healthcare domain has been open to public yet, thus the suggested model has not been designed for the healthcare domain only. The suggested model was trained with a wider domain data including the medical field. However, the subject or domain of the text does not greatly affect how and when to pause in Korean speech. The suggested prediction model can be used to implement the speech interfaces for medical and healthcare domain in a straightforward manner if the data from the required domain is provided.

Furthermore, to evaluate the proposed system practically, it should be adapted to generate standard Korean pronunciation. The performance comparison of pronunciation generation at the phrase unit with that of the *eo-jeol* unit will be a part of our continuing work.

Author Contributions: Conceptualization, M.K. and Y.J.; methodology, Y.J.; software, M.K.; validation, M.K., and Y.J.; formal analysis, M.K.; investigation, M.K.; resources, M.K. and Y.J.; data curation, Y.J.; writing—original draft preparation, M.K.; writing—review and editing, M.K. and Y.J; visualization, M.K., and Y.J.; supervision, H.-C.K.; project administration, H.-C.K.; funding acquisition, H.-C.K. Experimental data and several parts of results for this study are from PhD thesis of Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2013-2-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services).

Data Availability Statement: The news script data was generated from a Korean broadcast station which holds the copyright of the raw news scripts. Experimental data supporting the results of this study could be open to public when the copyright issue is resolved.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Box A1. Perceptual Prominence of Prosodic Labeling Systems.

정치인 # 자제들에 # 대한 # 병역 # 비리 # 수사가 # 시작된다고 # 하자 # 당장 # 선거를 # 앞둔 # 정치권은 # 긴장하고 # 있습니다///.
jeongchi-in # jajedeul-e # daehan # byeong-yeog # bili # susaga # sijagdoendago # haja # dangjang # seongeoleul # apdun # jeongchigwon-eun # ginjanghago # iss-seubnida///.
"Political circles facing elections ahead have their nerves on edge because police investigation into military draft scandals involving the children of the politicians has started."
#: no break,//: minor break,///: major break

Box A2. Variable Realization of Prosody, Dependent on the Speaker's Style.

[Announcer A]
부산지역 구, 군의회 의장단은///오늘//연제구의회에 모여
busanji-yeog gu, gun-uihoe uijangdan-eun///oneul//yeonjegu-uihoe-e mo-yeo
"Today, county and municipal councilors in Busan gathered in Yeonje parliament building."
[Announcer B]
이케다 일본 외무장관은//오늘///독도의 영유권에 관한 입장에는
ikeda ilbon oemujanggwan-eun//oneul///dogdo-ui yeong-yugwon-e gwanhan ibjang-eneun
"Today, Japanese Foreign Minister Ikeda (addressed) his position on the claim to Dokdo."

Box A3. Syntactic or Semantic Ambiguities.

고속버스가 중앙선을 침범해 마주오던 승용차를 들이받았습니다.
gosogbeoseuga jung-angseon-eul chimbeomhae maju-odeon seung-yongchaleul deul-ibad-ass-seubnida
a. 고속버스가//중앙선을 # 침범해///마주오던 # 승용차를//들이받았습니다.
Gosogbeoseuga//jung-angseon-eul # chimbeomhae///maju-odeon # seung-yongchaleul//deul-ibad-ass-seubnida
"An express bus drove over the center line and rammed into an oncoming car."
b. 고속버스가///중앙선을 # 침범해//마주오던 # 승용차를///들이받았습니다.
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."

References

1. Lim, S.G.; Jung, S.H.; Huh, J.H. Visual Algorithm of VR E-Sports for Online Health Care. *Healthcare* **2021**, *9*, 824. [CrossRef]
2. Kim, S.K.; Huh, J.H. Consistency of Medical Data Using Intelligent Neuron Faster R-CNN Algorithm for Smart Health Care Application. *Healthcare* **2020**, *8*, 185. [CrossRef]
3. Iqbal, N.; Ahmad, S.; Kim, D. Health Monitoring System for Elderly Patients Using Intelligent Task Mapping Mechanism in Closed Loop Healthcare Environment. *Symmetry* **2021**, *13*, 357. [CrossRef]
4. Kim, S.K.; Huh, J.H. Artificial Neural Network Blockchain Techniques for Healthcare System: Focusing on the Personal Health Records. *Electronics* **2020**, *9*, 763. [CrossRef]
5. Blackley, S.V.; Schubert, V.D.; Goss, F.R.; Al Assad, W.; Garabedian, P.M.; Zhou, L. Physician use of speech recognition versus typing in clinical documentation: A controlled observational study. *Int. J. Med. Inform.* **2020**, *141*, 104178. [CrossRef]
6. Wang, Y.; Jordan, C.S.; Laby, K.P.; Southard, J. Medical Tele-Robotic System with a Head Worn Device. U.S. Patent 7,262,573, 28 August 2007.
7. Amiribesheli, M.; Benmansour, A.; Bouchachia, A. A review of smart homes in healthcare. *J. Ambient. Intell. Humaniz. Comput.* **2015**, *6*, 495–517. [CrossRef]
8. Wolf, M.S.; Davis, T.C.; Shrank, W.; Rapp, D.N.; Bass, P.F.; Connor, U.M.; Clayman, M.; Parker, R.M. To err is human: Patient misinterpretations of prescription drug label instructions. *Patient Educ. Couns.* **2007**, *67*, 293–300. [CrossRef] [PubMed]
9. Wendt, O. *Assistive Technology: Principles and Applications for Communication Disorders and Special Education*; Brill: Leiden, The Netherlands, 2012.
10. Saz, O.; Yin, S.C.; Lleida, E.; Rose, R.; Vaquero, C.; Rodriguez, W.R. Tools and Technologies for Computer-Aided Speech and Language Therapy. *Speech Commun.* **2009**, *51*, 948–967. [CrossRef]
11. Selouani, S.A.; Yakoub, M.S.; O'Shaughnessy, D. Alternative Speech Communication System for Persons with Severe Speech Disorders. *EURASIP J. Adv. Signal Process.* **2009**, *2009*, 1–12. [CrossRef]
12. Potamianos, G.; Neti, C. Automatic speechreading of impaired speech. In Proceedings of the AVSP 2001-International Conference on Auditory-Visual Speech Processing, Yorktown Heights, NY, USA, 7–9 September 2001.
13. Cummins, N.; Scherer, S.; Krajewski, J.; Schnieder, S.; Epps, J.; Quatieri, T.F. A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* **2015**, *71*, 10–49. [CrossRef]
14. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Qadir, J.; Schuller, B.W. Deep representation learning in speech processing: Challenges, recent advances, and future trends. *arXiv* **2020**, arXiv:2001.00378.
15. Rana, R.; Latif, S.; Gururajan, R.; Gray, A.; Mackenzie, G.; Humphris, G.; Dunn, J. Automated screening for distress: A perspective for the future. *Eur. J. Cancer Care* **2019**, *28*, e13033. [CrossRef] [PubMed]
16. Taylor, P.; Black, A.W. Assigning phrase breaks from part-of-speech sequences. *Comput. Speech Lang.* **1998**, *12*, 99–117. [CrossRef]
17. Syrdal, A.K.; McGory, J. Inter-transcriber reliability of ToBI prosodic labeling. In Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China, 16–20 October 2000.
18. Hirschberg, J.; Prieto, P. Training intonational phrasing rules automatically for English and Spanish text-to-speech. *Speech Commun.* **1996**, *18*, 283–292. [CrossRef]

19. Lee, S.; Oh, Y.-H. Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems. *Speech Commun.* **1999**, *28*, 283–300. [CrossRef]
20. Fordyce, C.S. *Prosody Prediction for Speech Synthesis Using Transformational Rule-Based Learning*; Boston University: Boston, MA, USA, 1998.
21. Wightman, C.W.; Ostendorf, M. Automatic labeling of prosodic patterns. *IEEE Trans. Speech Audio Process.* **1994**, *2*, 469–481. [CrossRef]
22. Mittag, G.; Möller, S. Deep learning based assessment of synthetic speech naturalness. *arXiv* **2021**, arXiv:2104.11673.
23. Liu, J.M.; Xie, Z.Z.; Zhang, C.X.; Shi, G. A novel method for Mandarin speech synthesis by inserting prosodic structure prediction into Tacotron2. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 2809–2823. [CrossRef]
24. Yan, Y.; Jiang, J.; Yang, H. Mandarin Prosody Boundary Prediction based on Sequence-to-sequence Model. In Proceedings of the 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 12–14 June 2020; pp. 1013–1017.
25. Sloan, R.; Akhtar, S.S.; Li, B.; Shrivastava, R.; Gravano, A.; Hirschberg, J. Prosody prediction from syntactic, lexical, and word embedding features. In Proceedings of the 10th ISCA Speech Synthesis Workshop, Vienna, Austria, 20–22 September 2019; pp. 269–274.
26. Lu, C.; Zhang, P.; Yan, Y. Self-attention based prosodic boundary prediction for chinese speech synthesis. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7035–7039.
27. Du, Y.; Wu, Z.; Kang, S.; Su, D.; Yu, D.; Meng, H. Prosodic structure prediction using deep self-attention neural network. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 320–324.
28. Kocharov, D.; Kachkovskaia, T.; Skrelin, P. Prosodic boundary detection using syntactic and acoustic information. *Comput. Speech Lang.* **2019**, *53*, 231–241. [CrossRef]
29. Luo, L.; Yang, Z.; Yang, P.; Zhang, Y.; Wang, L.; Lin, H.; Wang, J. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics* **2018**, *34*, 1381–1388. [CrossRef]
30. Zheng, Y.; Tao, J.; Wen, Z.; Li, Y. BLSTM-CRF Based End-to-End Prosodic Boundary Prediction with Context Sensitive Embeddings in a Text-to-Speech Front-End. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 47–51. [CrossRef]
31. Jung, Y.; Cho, S.; Yoon, A.; Kwon, H.-C. Prediction of prosodic break using syntactic relations and prosodic features. In Proceedings of the Annual Conference on Human and Language Technology, Deagu, Korea, 12–13 October 2007; pp. 7–14.
32. Ostendorf, M.; Veilleux, N.M. A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Comput. Linguist.* **1994**, *20*, 27–54.
33. Mayo, C.; Aylett, M.; Ladd, D.R. Prosodic transcription of Glasgow English: An evaluation study of GlaToBI. In Proceedings of the Intonation: Theory, Models and Applications, Athens, Greece, 18–20 September 1997.
34. Jun, S.-A. Prosody in sentence processing: Korean vs. English. *UCLA Work. Pap. Phon.* **2005**, *104*, 26–45.
35. Jun, S.-A.; Lee, S.-H.; Kim, K.; Lee, Y.-J. Labeler agreement in transcribing korean intonation with K-toBI. In Proceedings of the INTERSPEECH 2000, Beijing, China, 16–20 October 2000; pp. 211–214.
36. Kim, S. *Rhythmic Units and Syntactic Structures in Korean: A Phonetic and Linguistic Study Aiming at Improving the Rhythmic Properties of Synthetic Speech*; Seoul National University: Seoul, Korea, 2002.
37. Kjelgaard, M.M.; Speer, S.R. Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *J. Mem. Lang.* **1999**, *40*, 153–194. [CrossRef]
38. Schafer, A.J. *Prosodic Parsing: The Role of Prosody in Sentence Comprehension*; University of Massachusetts Amherst: Amherst, MA, USA, 1997.
39. Lee, S.H. A Study of the Description System of Korean Prosodic Structure: K-ToBI Labeling System. *Linguistics* **2002**, *10*, 1–18.
40. Lee, S.; Oh, Y. The Modeling of Prosodic Phrasing and Pause Duration using CART. *Proc. KSCSP'98* **1998**, *15*, 81–86.
41. Wightman, C.W.; Syrdal, A.K.; Stemmer, G.; Conkie, A.; Beutnagel, M. Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis. *Group* **2000**, *1*, L3.
42. Kim, B.; Lee, G.G. Implementation of Korean TTS System based on Natural Language Processing. *Malsori* **2003**, *46*, 51–64.
43. Kwon, O.; Hong, M.-K.; Kang, S.-M.; Shin, J.-Y. AP, IP Prediction for Corpus-based Korean Text-to-speech. *Speech Sci.* **2002**, *9*, 25–34.
44. Sun, X.; Applebaum, T.H. Intonational phrase break prediction using decision tree and n-gram model. In Proceedings of the INTERSPEECH 2001, Aalborg, Denmark, 3–7 September 2001; pp. 537–540.
45. Jun, J.; Kim, H.; Kim, D.; Lee, Y. Prosodic-Boundary Prediction for Korean TTS System. In Proceedings of the Acoustical Society of Korea, Busan, Korea, 10–11 October 2002; pp. 77–82.
46. Kim, S.; Kim, B.; Jeong, M.; Lee, G.G. Using CRF to Predict Phrase Breaks in Korean. In Proceedings of the 17th Annual Conference on Human and Cognitive Language Technology, Seoul, Koea, 21–22 October 2005; pp. 134–138.
47. Maragoudakis, M.; Zervas, P.; Fakotakis, N.; Kokkinakis, G. A data-driven framework for intonational phrase break prediction. In Proceedings of the International Conference on Text, Speech and Dialogue, České Budéjovice, Czech Republic, 8–12 September 2003; pp. 189–197. [CrossRef]

48. Viana, M.C.; Oliveira, L.C.; Mata, A.I. Prosodic phrasing: Machine and human evaluation. *Int. J. Speech Technol.* **2003**, *6*, 83–94. [CrossRef]
49. Grice, M.; Reyelt, M.; Benzmuller, R.; Mayer, J.; Batliner, A. Consistency in transcription and labelling of German intonation with GToBI. In Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP 1996), Philadelphia, PA, USA, 3–6 October 1996; pp. 1716–1719. [CrossRef]
50. Pitrelli, J.F.; Beckman, M.E.; Hirschberg, J. Evaluation of prosodic transcription labeling reliability in the tobi framework. In Proceedings of the 3rd International Conference on Spoken Language Processing, Yokohama, Japan, 18–22 September 1994; pp. 123–126.
51. Grawe, P.H. *Nonparametric Statistics for the Behavioral Sciences*; McGraw-Hill: New York, NY, USA, 1988.
52. Carletta, J. Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.* **1996**, *22*, 249–254.
53. Kang, M.-y.; Jung, S.-W.; Park, K.-s.; Kwon, H.-C. Part-of-speech tagging using word probability based on category patterns. In Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico, 18–24 February 2007; pp. 119–130.
54. Manning, C.; Schutze, H. *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, USA, 1999.
55. Jung, Y.; Yoon, A.; Kwon, H.-C. Grapheme-to-phoneme conversion of Arabic numeral expressions for embedded TTS systems. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *15*, 296–309. [CrossRef]
56. Yarowsky, D. Homograph disambiguation in text-to-speech synthesis. In *Progress in Speech Synthesis*; Springer: Berlin/Heidelberg, Germany, 1997; pp. 157–172.
57. Chen, T.; Xu, R.F.; He, Y.L.; Wang, X. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Syst. Appl.* **2017**, *72*, 221–230. [CrossRef]