Automatic Corpus-based Tone Prediction using K-TOBI Representation *

Jin-seok Lee, Byeongchang Kim and **Gary Geunbae Lee** †
Department of Computer Science & Engineering
Pohang University of Science & Technology
Pohang, 790-784, South Korea
{likeoxy, bckim, gblee}@postech.ac.kr

Abstract

In this paper, we present a prosody generation architecture based on K-ToBI (Korean Tone and Break Index) representation.

ToBI is a multi-tier representation system based on linguistic knowledge to transcribe events in an utterance. The TTS system which adopts ToBI as an intermediate representation is known to exhibit higher flexibility, modularity and domain/task portability compared with the direct prosody generation TTS systems. However, the cost of corpus preparation is very expensive for practical-level performance because the ToBI labeled corpus has been manually constructed by many prosody experts and normally requires large amount of data for statistical prosody modeling.

Contrary to previous ToBI-based systems, this paper proposes a new method which transcribes the K-ToBI labels completely automatically in Korean speech. We developed automatic corpus-based K-TOBI labeling tools and prediction methods based on several lexico-syntactic linguistic features for decision-tree induction. We demonstrated the performance of F0 generation from automatically predicted K-ToBI labels, and confirmed that the performance is reasonably comparable with state-of-the-art direct prosody generation methods and previous TOBI-based methods.

1 Introduction

For the past few years, a great deal of interests have arisen in high quality text-to-speech (TTS) systems (van Santen et al., 1997). One of the essential problems in developing high quality TTS systems is to generate prosody from texts. The prosody in the TTS system increases intelligibility and naturalness of synthesized speech. In general, prosody refers to certain properties of the speech signal, such as audible changes in pitch, loudness and syllable length. Its physical features are related to prosodic phrasing, fundamental frequency, segmental duration and energy of synthesized speech. In TTS systems especially, prosodic phrasing and fundamental frequency contour generation are the most important tasks to produce natural speech.

In order to produce high quality speech, several methods have been devised so far. We can classify fundamental frequency (F0) contour generation according to level of representation in prosodic phenomena; that is, acoustic level, perceptual level and linguistic level (Dutoit, 1997). This paper is related to a model using linguistic knowledge. Specifically, we deal with prosodic phrasing and F0 contour generation among several prosodic features using K-ToBI labeling system. K-ToBI (Beckman and Jun, 1998) is a Korean implementation of the standard ToBI system.

Our approach has the following specific design goals:

1) The changes of intonation in Korean language are very weak, so the K-ToBI labeling system must be simpler than the other language ToBI systems.

2)K-ToBI, as an intermediate representation, normally increases system-level modularity, flexibility and domain/task portability, but should be implemented with no performance degradation.

3) Labeling ToBI on speech corpus is normally very laborious and time-consuming. However, large portion of Korean speech is mainly monotonous in intonation. Therefore, we can

* This research was partly supported by Hyundai Autonet, Inc. and by BK21 program of Korea Ministry of Education

† Currently visiting CSLI Stanford University

do automatic tone labeling by linguistic corpus transformation with simple lexico-syntactic rules.

In the next section, previous researches of K-ToBI representation and automatic tone labeling technique are discussed with some features of Korean prosody modeling in contrast to that of English. Section 3 explains our tone label prediction method including automatic K-ToBI labeling from large speech corpus. Section 4 shows extensive experimental results and comparisons to demonstrate the performance of the system, and section 5 draws some conclusions.

2 Previous Researches

2.1 K-ToBI system

ToBI (for the Tones and Break Indices) is a system to transcribe intonation patterns and other aspects of the prosody of English utterances (Silverman et al., 1992). K-ToBI (Korean ToBI) is a prosodic transcription convention for standard (Seoul) Korean (Jun, 2000).

The ToBI system consists of multiple tiers, and each label represents the prosodic events of utterances. These tiers consist of a tone tier for specifying the intonation pattern, an orthographic tier for specifying the words and their boundaries, a break-index tier for marking the prosodic grouping of words, and a miscellaneous tier for annotating comments that can be used to mark events, such as a cough, laughter, etc.

The K-ToBI labeling system is similar to English ToBI in the view of principal. The K-ToBI labeling system includes a word tier, a phonological tone tier and phonetic tone tier, a break-index tier, and a miscellaneous tier. The intonational structure of Korean is defined as shown in Figure 1 (Jun, 2000).

This structure has two intonationally defined prosodic units: Intonation Phrase (IP) and Accentual Phrase (AP). An AP is smaller than an IP and larger than a phonological word which is a lexical item plus case markers or postpositions. Therefore, an utterance consists of one or more IPs, and an IP consists of one or more APs. An AP is marked by a phrasal tone such as THLH (T means initial tone, that is, H or L). Similarly, an IP is marked by a boundary tone (%) and final lengthening such as L%, H%, LH%, etc.

The Korean language has no strong accents,

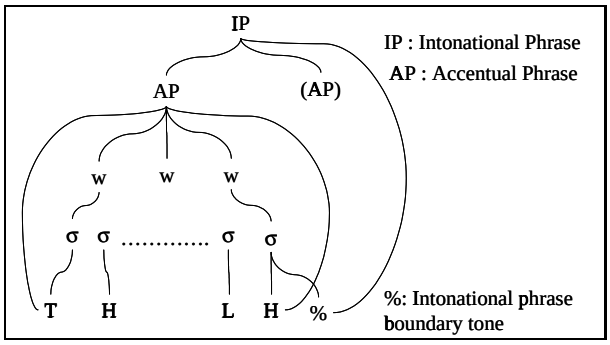

Figure 1: Intonational structure of standard Korean. (IP : Intonation Phrase, AP : Accentual Phrase, w : phonological word, σ : syllable, T: Initial tone, % : Intonation phrase boundary tone)

in contrast to English or Japanese languages, so a tone tier at K-ToBI can be relatively simple and the symbol corresponding to the tonal element just represents changes of intonation. The current version of K-ToBI expands a tone tier into two levels: a phonological tone tier and a phonetic tone tier, in order to describe the surface tonal patterns that are not easily predictable from the distinctive underlying tones. A word tier in K-ToBI corresponds to the 'orthographic tier' in English ToBI. What constitutes a 'word' in Korean is a matter of some debate, but current version considers the 'word' as an eojeol[1] which consists of one or more morphemes. A break-index tier in K-ToBI represents 4 steps of phrase breaks compared to 5 steps in English ToBI, and is marked after all words that have been transcribed in the word tier. The break indices represent a rating for the degree of juncture perceived between each pair of words and also between the final word and the silence at the end of utterance.

2.2 ToBI-represented prosody modeling

Many of ToBI-related prosody modeling researches have been conducted for ToBI label prediction and fundamental frequency contour generation from ToBI labels. (Black and Hunt, 1996) predicts fundamental frequency contours using linear regression functions from manually constructed ToBI-labeled corpus. They

[1] Eojeol is a segmental unit divided by white space in written text

obtained RMSE (Root Mean Squared Error) of 34.8Hz and 0.62 correlation coefficient when compared with originals on the test data, which is a significant improvement compared to previous rule-driven methods. They applied the same technique to J-ToBI (Japanese ToBI) and obtained RMSE of 20.9Hz and 0.70 correlation coefficient (Black and Hunt, 1996). However, these methods require massive amounts of manually constructed ToBI labeled corpus. In this paper we propose an automatic K-ToBI tone label prediction method without using manually constructed ToBI corpus while achieving same level of performance.

(Ross, 1995) proposed a model that predicts the ToBI label and fundamental frequency using CART-based probability tree with Markov sequence assumption. The method predicted ToBI labels including accent location, symbolic tones and relative prominence level from a text tagged with part-of-speech (POS) labels and marked for the prosodic constituent structure, and showed an accuracy of prediction about 80% at the pitch accent type and about 67% at the boundary tone type. For fundamental frequency and energy generation, they obtained RMSE of 34.74% and 3.48% respectively. The merit of this approach was that the parameters can be automatically estimated from labeled speech, but the system only predicted the abstract prosodic label instead of the original ToBI label. Furthermore, because the training data was so small, the results could not be extended to general cases. We will predict tone labels using not only linguistic knowledge but also pitch values extracted from large amount of speech and transcription files.

(Lee et al., 1998) proposed a method of K-ToBI labeling from POS tagged texts and generated an fundamental frequency contour from the K-ToBI-labeled corpus using the method similar to (Black and Hunt, 1996). They showed accuracy of 77.2% for IP boundary and 72% for AP boundary prediction, and obtained RMSE of 23.5Hz and 0.55 correlation coefficient for fundamental frequency contour generation.

(Lee, 2000) proposed a tree-based prosody modeling by appling bootstrap aggregating and born again tree techniques proposed in the statistics community to predict the prosodic components more accurately. He applied not only morphological knowledge but also syntax-based features to predict K-ToBI tone label and generating fundamental frequency contour. The performance was RMSE of 35.65Hz and 0.729 correlation coefficient.

However all the previous approaches still suffer from data sparseness problem due to manual labeling of prosody data. We try to avoid such problem by using a more general and simplified method of prosodic phrasing and fundamental frequency generation using a large sized, but automatically K-ToBI-labeled corpus.

3 Tone Label Prediction using K-ToBI

3.1 Automatic K-ToBI tone labeling

Because K-ToBI tone implies a variety of linguistic knowledge and changes in intonation, such as the fundamental frequency contour, it is an important part of TTS systems. However, constructing the K-ToBI labeled corpora is very difficult and time consuming because they are usually constructed by some experts who have phonological knowledge by referring to pitch contours while hearing a speech corpus. In addition, such manually constructed corpora can lack consistency due to the differences between annotators. However, predicting K-ToBI tone labels using a statistical method inevitably requires a large corpus. Alternatively, in this section, we propose an automatic method of building and refining large corpus for K-ToBI labeling. Figure 2 shows the whole flow of our automatic K-ToBI tone labeling process.

First, we extract pitches from a speech data and align the speech with its phone scripts that have been manually corrected. Second, we align the phone scripts with the extracted pitches. Third, the phone sequence that has been aligned with the pitches is segmented into phrases by comparing with the phrase indices. Using the above sequences, we can obtain a corpus that was aligned with time and segmented into prosodic phrases. Finally, we calculate the max, min and average values of pitch sequences in each prosodic phrase, which would become criteria that assign the 'L' tone or the 'H' tone. Since humans have a limited perception of changes in intonation, when we specify ToBI symbols into syllable or IP boundary, ambiguous regions naturally occur. We specify the

ambiguous regions into 'M' tone which consists of 'M+' and 'M-' again. The 'M+' tone is closer to the region of the 'H' tone and the 'M-' tone is the opposite.

As shown in Equation 1, the region of the 'M' tone which we cannot directly determine by only reading the values is larger than that of the 'H' tone or the 'L' tone. We assign these ambiguous tones to the 'M' tone first, and then re-map the 'M' tone to 'H' or 'L' tone according to the context information.

$$X = \begin{cases} L & (IP_{min} \leq X < L_{upper}) \\ M- & (L_{upper} \leq X < IP_{avg}) \\ M+ & (IP_{avg} \leq X \leq H_{lower}) \\ H & (H_{lower} < X \leq IP_{max}) \\ \text{Not apply} & (X = 0) \end{cases}$$

IP_{min}: the minimum pitch range value in the IP
IP_{max}: the maximum pitch range value in the IP
IP_{avg}: the average pitch range value in the IP
L_{upper}: the mean of the minimum pitch value and the average pitch value in the IP
H_{lower}: the mean of the maximum pitch value and the average pitch value in the IP

(1)

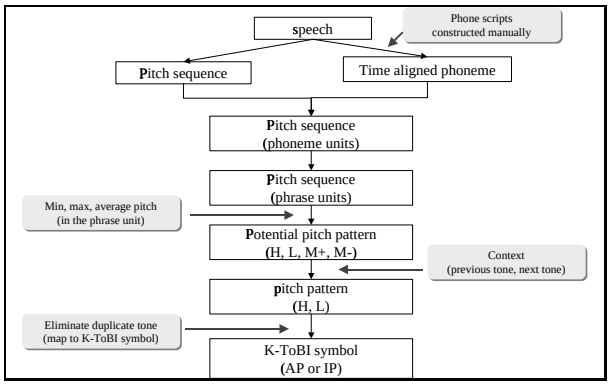

Figure 2: Automatic K-ToBI tone labeling process

3.2 K-ToBI IP tone label prediction

Even though the change of intonation is light in Korean, the change near the intonational phrase boundary is more active than others. Therefore, we must achieve reasonable quality of predictions for these boundary tones. We predict the boundary tones using the state-of-the-art tree-base method, which is a general technique in tone prediction, and is simple for training and testing. Furthermore, the tree-based method is useful in cases where we select just one category out of many categories. The K-ToBI has L%, H%, LH%, HL%, LHL% and HLH% as boundary tones. These tone symbols are labeled automatically by the method explained in section 3.1, and we obtain 23,725 items in 10,149 sentences as a result. We set 20,933 items for training and 2,792 items for test, respectively. Even the tree-based method is well-known, the feature selection is still problematic since the performance is dominated by the selected features. The followings are our choice of features for Korean IP tone prediction.

- LeftPOS, RightsPOS, NextLeftPOS: the leftmost POS (part-of-speech) and rightmost POS of final Eojeol in prosodic phrase, and leftmost POS of next Eojeol. In agglutinative languages such as Korean, POS information of morphemes has strong influence to the intonation.
- SylStr: the onset, nucleus, coda of pronunciation at the final syllable. In Korean, the syllable that started as a plosive sound or glottal sound tends to be pronounced with high tone.
- SylType: the configuration type of final syllable in a prosodic phrase. This feature shows the configuration of vowel and consonant in a syllable.
- PuncType: the punctuation mark type. The syllable that is followed by a question mark or an exclamation mark tends to be pronounced with high tone.
- Phrloc: the location of the prosodic phrase. Usually, the final phrase in the sentence tends to be pronounced with low tone. We assigned 'Initial', 'Middle' and 'Final' to this feature.
- Phrlen: the length of the prosodic phrase in a syllable. The intonation changes according to the length of the phrase.
- RelPos: the relative position of a prosodic phrase in utterance. The relative location in prosodic phrases has some influence to the intonation in Korean.

3.3 K-ToBI AP tone label prediction

The AP tone prediction plays a role of smoothing between prosodic phrases. As shown in

Figure 1, the non-boundary tones in Korean prosody are assigned to each syllable. Therefore, we predict the K-ToBI tone label as 'H' tone or 'L' tone per syllable. We obtained 265,726 items of L tone and 121,322 items of 'H' tone automatically from the K-ToBI tone labeled corpus. We also used the same tree-based method as IP tone prediction, but features for the decision-tree are newly designed for best AP label prediction as follows:

- NucleusPOS, CodaPOS: the POS of the nucleus and coda respectively. Usually, the POS of onset is the same as that of the nucleus in Korean. Therefore, only two POS's are used. The intonation of syllable is presumed to be related to the POS of the syllable.
- SylType: the configuration type of each syllable.
- LocinPhr: location of the current syllable in a prosodic phrase. Usually, changes in the 'H' or 'L' tone are more active at the second syllable in the prosodic phrase. We assigned 'First', 'Second' and 'Others' to this feature.
- LengthPhr: the length of the current prosodic phrase. The short prosodic phrase tends to be pronounced with 'H' tone.
- SylPhrb, SylPhre: the distance from the current syllable to the beginning or from the end of current prosodic phrase. If some absolute patterns exist in a prosodic phrase, these features have some influence on the decision of tones.
- RatioSylPhrb: the ratio of SylPhrb over the length of the current phrase. This feature is useful for the decision of tones, if there is a regular pattern in a prosodic phrase.

3.4 Fundamental frequency contour generation

In most synthesizers, the task of generating a prosodic tone using ToBI label system consists of two sub-tasks: the prediction of intonation labels (AP tone and IP boundary tone in ToBI) from input text (previously described), and the generation of a fundamental frequency contour from those labels and other information. We used popular linear regression method to generate fundamental frequency from K-ToBI labels (Black and Hunt, 1996). This method does not require any other rules for label types, and is general enough for many other languages. Our prediction formula is as follows:

$$target = w_1 f_1 + w_2 f_2 + \cdots + w_n f_n + I$$

where f_i are the features that contribute to the fundamental frequency, and we can decide the weight $w_1 \sim w_n$ and I through simple linear regression. We applied the above formula to every syllable and obtained a target value of the fundamental frequency. We prepared pitch values extracted from a speech file, divided them into five sections, and predicted the fundamental frequency at every point. The following features overlap somewhat with features we developed for boundary and non-boundary tone prediction, except some more features that, we anticipate, would generate pitches.

- SylStr: onset, nucleus, coda of pronunciation form at current syllable.
- LocinPhr: the location of the current syllable in the prosodic phrase.
- Phrbndry: the type of boundary tone at the preceding phrase. The pitch of the current syllable is affected by the preceding phrase boundary tone type.
- Toneseq: three preceding tones, one current tone, and three following tones. If some patterns of intonation of the prosodic phrase exist, these features are useful.
- LengthPhr: the length of current prosodic phrase.
- SylPhrb, SylPhre: the distance from the current syllable to the beginning or to the end of current prosodic phrases.
- RatioSylPhrb: the ratio of SylPhrb over the length of the current phrase.

4 Experimental Results

4.1 Corpus analysis

The experiments are performed on a Korean news story database, called MBCNEWSDB, of spoken Korean directly recorded from broadcast

news programs. The size of the database is now 10,149 sentences (122,047 words) and is continuously growing. For our prosody prediction experiments, the database was POS tagged and automatically ToBI-labeled, as described in section 3.1.

The occurrence frequencies of the tones are shown in Table 1, which shows weak tonal changes in Korean. The 'L%' tone in IP boundary dominates other boundary tones because the corpus is a news broadcast corpus.

Figure 3 through Figure 5 show some statistics of the corpus: the number of eojeols per sentence, the number of IPs per sentence and the number of APs per IP. The 'eojeol' means 'word' in the K-ToBI, which corresponds to the orthographic boundary in English ToBI.

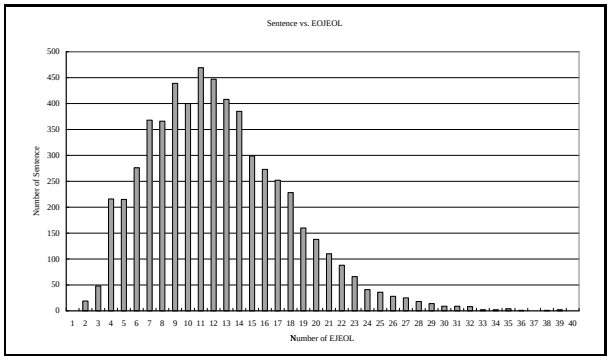

Figure 3: The number of eojeols per sentence.

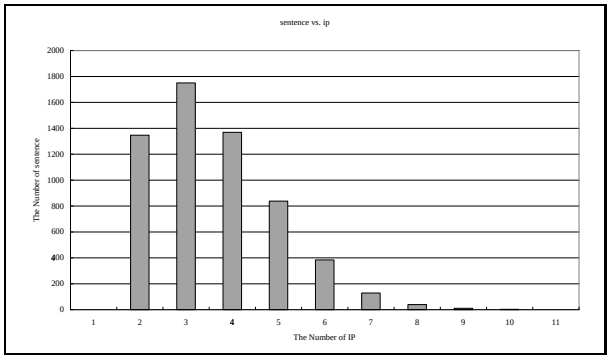

Figure 4: The number of prosodic phrases (IPs) per sentence.

4.2 Tone label prediction

In K-ToBI, a boundary tone has L%, H%, LH%, HL%, LHL% and HLH% values. We obtained 23,725 items in 10,149 MBCNEWSDB sentences, and used 8780 sentences for training

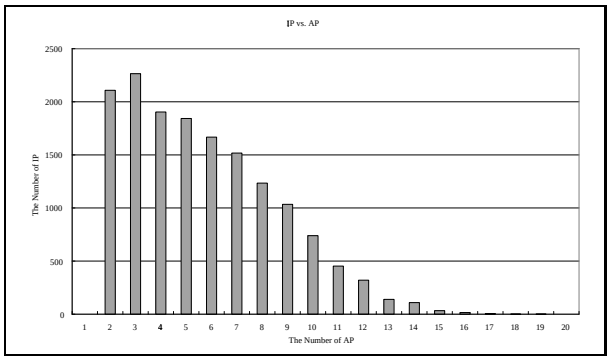

Figure 5: The number of accentual phrases per prosodic phrase.

and 1369 sentences for test. Table 2 shows the configuration of each tone in the corpus. Because the corpus comes from TV broadcasted news program, most of the intonation is flat and the 'L%' tone is dominant. In addition, since most of the sentences are declarative sentences, the pitch generally declines as the time passes.

We use the decision tree after pruning since unpruned tree could be overfited for this kind of skewed training data. We set the pruning confidence level to 25% (Quinlan, 1983). Table 4 shows IP tone prediction results. The decision criterion at the root node of the decision tree is the length of the prosodic phrase. That is, when the value 'Phrlen' feature is less than 10 (in syllables), the boundary tone becomes 'L%' tone. The next decision is based on the location of the prosodic phrase (Phrloc feature). If the 'Phrloc' feature is 'F' (if this phrase is located at Final in the sentence), the tree decided the tone as 'L%' at that time. Next, the tree inspects 'Coda', 'Nucleus' and 'Pelpos' features in that order.

For the AP tone label prediction, we obtained 387,048 items as non-boundary tones in the automatic K-ToBI corpus. This tone consists of only 'L' and 'H' tones. Table 4 shows the number of total non-boundary tones. As shown in the table, the 'L' tone exists more than 'H' tone. We used 10 features from 6 categories and 4 continuous values, as shown in section 3.3, for decision-tree training. The result of our prediction is shown in Table 5.

In K-ToBI, accentual phrase represents both the phonological tone tier and the phonetic tone tier. The phonological tone tier represents the underlying tones which are not directly pre-

Table 1: Occurrence frequency of tones in the corpus (per syllable)

L	H	L%	H%	LH%	HL%	LHL%	HLH%
265726	121322	18914	1436	1522	1161	593	99

Table 2: The configuration of each boundary tone in the MBCNEWSDB corpus.

	L%	H%	LH%	HL%	LHL%	HLH%	TOTAL
TRAIN	16567	1322	1387	1026	540	91	20933
TEST	2347	114	135	135	53	8	2792
TOTAL	18914	1436	1522	1161	593	99	23725

dictable from the surface patterns, and consists of a rising contour 'LHa' and ending IP tone 'T%'. The phonetic tone tier represents a surface tonal pattern and is principally assigned with an 'L' or 'H' tone for each syllable. Therefore, this result shows prediction of surface tonal pattern in the phonetic tone tier. Through this experiment, we obtained total 70% of accuracy. In the tree, the first node makes a decision if the ratio of the distance from beginning of the current prosodic phrase to the current syllable over the length of the current prosodic phrase ('RatioSylPhrb' feature) is larger than 0.2.

4.3 Fundamental frequency generation

We predicted five pitch values for each syllable and generate fundamental frequency with interpolated methods based on the predicted values. We detected pitch values per 16ms from speech with SFS tool (Huckvale, 1996). For training and test, we extracted a feature set from an automatically created file that is aligned with pitch sequences and phone sequences. Further, we eliminate items with '0' for pitch value as a noise for the constructing model.

As described before, we used ten category and four real value features to construct a linear regression model. Since we extracted five pitches from each syllable, we construct a linear regression model for each pitch, and obtained a total of five models. Since the predicted response is a vector, we compute the RMSE and correlation coefficients for each element of the vectors and averaged them. Table 6 shows the result of our generation model with ToBI and its comparision with direct generation without ToBI representation.

The table shows that ToBI-based generation is surely competitive to the direct generation. The result suggests that we can overcome the performance deficiency caused by ToBI with automatically labeled large corpus for training especially for light tonal change languages, such as Korean.

To compare with previous ToBI-based results, we used a relative measure called 'Relative-MSE (mean squared error)', dividing the previous RMSE by the standard deviation of corpus. This relative measure can overcome the differences between the features of the corpus used in the model. According to the table, (Black and Hunt, 1996)'s original results of 34.8Hz RMSE using linear regression model is worse than our 22.2 Hz of RMSE results. According to the new measure, (Ross and Ostendorf, 1999) obtained an RMSE of 34.74Hz, with the standard deviation of corpus of 42.8Hz, which gives 0.66 Relative-MSE similar to ours. But their corpus was as short as 48 minutes compared with our 1100 minutes.

For the Korean language, (Lee, 2000) obtained an RMSE of 35.65Hz and the deviation of the corpus is 50.74Hz. He obtained 0.49 Relative-MSE, which was better than our 0.66. However, he used a small manually transcribed K-ToBI corpus compared with our large automatically labeled corpus.

5 Conclusion

This paper presents a prosody generation method based on the K-ToBI system for Korean. Our main contributions include adopting

Table 3: The result of intonation boundary tone prediction (pruning: confidence level = 25%).

		Train	Test
Error rate	Before pruning	16.00%	19.60%
	After pruning	17.90%	16.10%

Table 4: The configuration of each non-boundary tone used in the training and the test.

	L	H	TOTAL
TRAIN	230650	105033	335683
TEST	35076	16289	51365
TOTAL	265726	121322	387048

the K-ToBI labels as an interim representation of the prosody without performance degradation. As a result, the architecture of the TTS system can have more flexibility and portability by using the ToBI system, and the TTS system can also be modularized. Another contribution involves the automatic ToBI tone labeling techniques used to construct a large size consistent training and test corpora. Therefore, we solved the problem that a ToBI-labeled corpus must be manually constructed. As shown in our results, the performance of prosody generation using the K-ToBI system is competitive to those without the ToBI system, and also competitive to the previous state-of-the-art prosody generation researches.

In the future, we have to automatically construct dialog style ToBI labeled corpus to verify our prediction model more thoroughly since current MBCNEWSDB has biased tone labels because it is a reading corpus for broadcast news scripts. A new algorithm from machine learning community especially to handle skewed and biased training data will also be greatly helpful to enhance the prediciton accuracy.

References

M.E. Beckman and S. Jun. 1998. K-ToBI (KOREAN ToBI) labeling convention. In *Proceedings of the study of Korean prosody.*

A. W. Black and A.J. Hunt. 1996. Generating f0 contours from ToBI lables using linear regression. In *Proceedings of the international conference on spoken language processing(ICSLP)*, pages 1385–1388.

C. d'Alessandro and P. Mertens. 1995. Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language*, 5(3):257–288.

T. Dutoit. 1997. *An introducyion to Text-to-Speech synthesis.* Kluwer Academic Publishers.

H. Fujisaki and S. Ohno. 1995. Analysis and modeling of fundamental frequency contours of english utterances. In *Proceedings of the EUROSPEECH '95*, pages 985–988.

M. Huckvale, 1996. *Speech Filing System*, sfs release 3 edition, March.

S. Jun. 2000. K-ToBI(KOREAN ToBI) labelling conventions (version 3.0, revised in january 2000). In *Proceedings of The Phonetic Society of Korea Workshop*, pages 105–140.

Y.J. Lee, S.H. Lee, J.J. Kim, H.J. Ko, Y.I. Kim, S.H. Kim, and J.C. Lee. 1998. A computational algorithm for f0 contour generation in korean developed with prosodically labeled databases using K-ToBI system. In *Proceedings of the international conference on spoken language processing(ICSLP)*, pages 1995–1998.

S. Lee. 2000. *Tree-based Modeling of Prosody for Korean TTS system.* Ph.D. thesis, Korea Advanced Institute of Science and Technology.

G. Mohler and A. Conkie. 1998. Parametric modeling of intonation using vector quantization. In *Third Speech Synthesis Workshop.*

J. R. Quinlan. 1983. *C4.5: Programs for Ma-*

Table 5: The result of non-boundary tone prediction.

		Train	Test
Error rate	Before pruning	30.00%	30.20%
	After pruning	30.20%	30.40%

Table 6: The result of fundamental frequency generation.

	Using ToBI	Without ToBI
RMSE	22.219	22.157
Corr	0.595	0.598
Relative-MSE	0.660	0.657

chine Learning. Morgan Kaufmann.

K.N. Ross and M. Ostendorf. 1999. A dynamical system model for generating fundamental frequency for speech synthesis. *IEEE Transaction on Speech and Audio Processing*, 7(3):259–309.

K.N. Ross. 1995. *Modeling of Intonation for Speech Synthesis*. Ph.D. thesis, Boston University College of Engineering.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. ToBI: A standard for labeling english prosody. In *Proceedings of the international conference on spoken language processing(ICSLP)*, pages 867–870.

P. Taylor. 1995. The rise/fall/connection model of intonation. *Speech Communication*, 15:169–186.

Jan P.H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg. 1997. *Progress in Speech Synthesis*. Springer-Verlag.