Modality-Based Sentence-Final Intonation Prediction for Korean Conversational-Style Text-to-Speech Systems

Seung-Shin Oh and Sang-Hun Kim

ABSTRACT—This letter presents a prediction model for sentence-final intonations for Korean conversational-style text-to-speech systems in which we introduce the linguistic feature of 'modality' as a new parameter. Based on their function and meaning, we classify tonal forms in speech data into tone types meaningful for speech synthesis and use the result of this classification to build our prediction model using a tree structured classification algorithm. In order to show that modality is more effective for the prediction model than features such as sentence type or speech act, an experiment is performed on a test set of 970 utterances with a training set of 3,883 utterances. The results show that modality makes a higher contribution to the determination of sentence-final intonation than sentence type or speech act, and that prediction accuracy improves up to 25% when the feature of modality is introduced.

Keywords—Sentence-final intonation, prediction model, modality, text-to-speech system, conversational-style.

I. Introduction

Among all the prosodic phenomena in Korean spontaneous speech, sentence-final intonation, which is realized as a tonal pattern on the last syllable, is the most determinant feature indicating the speaker's attitude in conveying a message. Unlike read speech, spontaneous speech shows a variety of sentence-final intonation patterns.

In conversational-style speech synthesis, the generation of sentence-final intonation raises two issues. One is the classification and listing of all the meaningful tone types and the other is the extraction of the most effective linguistic feature for the prediction of tone types from conversational text. This letter aims to tackle both these issues.

In making a glossary of tone types for sentence-final positions, we find that the classification systems proposed in previous research [1], [2] are not refined enough for conversational-style speech. We therefore propose a classification of sentence-final intonations for conversational speech based on their pragmatic function and meaning.

In selecting the linguistic features as parameters for prediction modeling in previous studies, it was common to assume that the main features related to intonation patterns were sentence type and speech act. In this letter, however, we propose *modality* as a new parameter. Modality is a semantic category defined as "the speaker's (subjective) attitude or opinion towards the content of a sentence" [3], while *sentence type* is a syntactic category and *speech act* is a pragmatic classification. The contribution of these three features to prediction performance was evaluated by an experiment in order to select the most effective one for prediction modeling of sentence final intonations.

II. Classification of Sentence-Final Intonation

1. Data

The speech data used for this analysis was created by recording a set of sentences which were extracted from a large corpus of 970,000 conversational style sentences in different domains. The set was selected so that it contained all the sentence ending types occurring in the corpus. The resulting

Manuscript received June 21, 2006; revised Aug. 29, 2006.

This research was supported by the Ministry of Information and Communication of the Republic of Korea.

Seung-Shin Oh (phone: +82 42 860 1360, email: oss63354@etri.re.kr) and Sang-Hun Kim (email: ksh@etri.re.kr) are with Embedded Software Research Division, ETRI, Daejeon, Korea.

corpus consisting of 4,853 sentences covering 365 sentence endings was recorded by a female radio announcer.

2. Classification of Tone Types

Different annotation systems have been proposed in the past regarding the sentence-final tones in Korean, such as K-ToBI (Korean Tone and Break Indices) [1] and the system proposed by H.Y. Lee [2]. The classification of IP boundary tones differs in these two systems in that the former classifies tone types just by their shapes, while the latter also takes the height of pitch into consideration by introducing mid (M) between the low (L) and high (H) tones. This will be our approach as well, since distinction between M and H is necessary for speech synthesis: an H (or LH) tone has the potential to turn a sentence into a question while an M (or LM) tone has no such functionality. However, we have further refined the classification of sentence-final tone types proposed in [2] with regard to L%, HL%, and LHL%. This is shown in Table 1.[1)]

We have first divided the low-level tones into two types: an ordinary low-level or gently falling tone and a low, rapidly falling tone. A low-level or gently falling tone (Fig. 1) indicates the speaker's certainty or confidence while a rapidly falling low tone (Fig. 2) does not express such an attitude.[2)]

The rapidly falling tone tends to occur in a different context

Table 1. Proposed sentence-final tone type classification.

K-ToBI	Lee [2]	Our proposal
L%	L%	LL%
		L%
H%	M%	M%
	H%	H%
LH%	LM%	LM%
	LH%	LH%
HL%	ML%	ML%
	HL%	HL%
		TL%
LHL%	LHL%	LML%
		LHL%
HLH%	HLH%	HLH%
LHLH%	-	-
HLHL%	-	-
HLHLH%	-	-

1) The glossary of tone types may be expanded in the light of new data.

2) The Korean alphabet in this letter is romanized according to the Revised Romanization system.

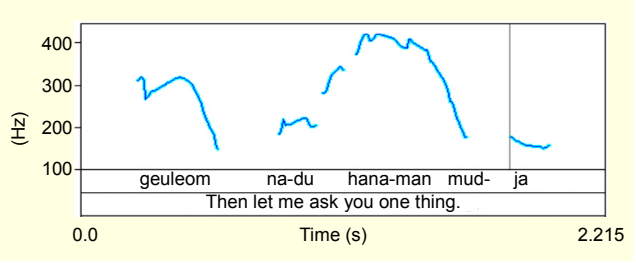

Fig. 1. An example of L%.

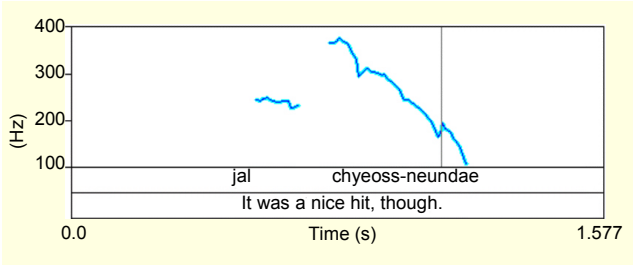

Fig. 2. An example of LL%

than the ordinary low-level tone, such as in a monolog or when reading noun phrases such as titles that don't need to show any attitude of the speaker. We distinguish this tone pattern from the normal low-level tone (L%) and consider it unmarked in terms of speaker's attitudinal meaning. We assign it the new symbol LL%. We have divided the high fall tone (HL%) into two levels as well. In Fig. 3, the maximum fundamental frequency of the ending tone is above the middle of the speaker's pitch range for that sentence, but not as high as the one in Fig. 4, which reaches the top of the speaker's pitch range. This difference affects the function of the sentence. The higher tone turns the

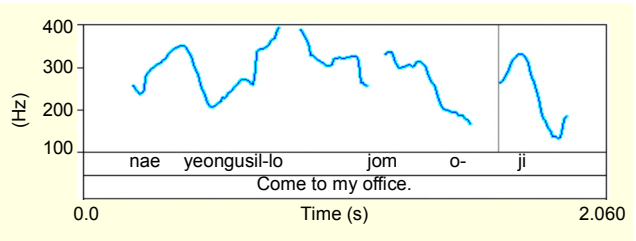

Fig. 3. An example of HL%.

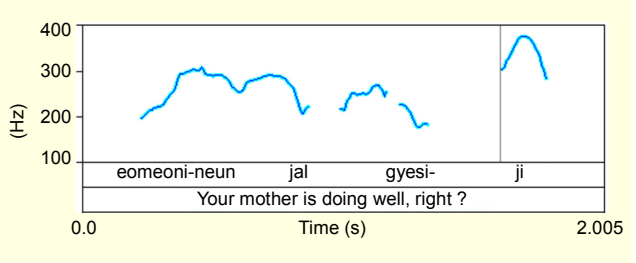

Fig. 4. An example of TL%.

sentence into a question whereas the lower tone just expresses the speaker's urge for the listener's acknowledgment. This tone is assigned the symbol TL% to distinguish it from HL%.

Finally, in addition to LHL% tones, we introduced LML% tones since LHL% can be used to express a stronger attitude of exclamation or persuasion than LML%.

a) Geuleoh-gun-yo./LML%: statement (or weak exclamation)
"I see." (I understand.)
b) Ah, geulaess-eoss-guna!/LHL% : exclamation
"Ah, now I understand!"

The HLH% tone indicating an attitude of objection was observed in the data only once. No MLM% tone was observed nor introduced in our final classification.

III. Prediction of Sentence-Final Intonations

1. Linguistic Features as Candidate Parameters

Until now, the most probable parameter candidates for intonation prediction were sentence type and speech act. The relation between sentence type and intonation has often been discussed [2], [4], [5], although it has also been pointed out that sentence type does not necessarily correspond to intonation type [4]. Speech act information is yet another feature which is regarded as potentially relevant to intonation type [4].

We now propose to introduce the modality of a sentence as a parameter for prediction modeling. In Korean, modality is mostly expressed by sentence endings (and additionally by modal verbs or modal adverbs) and can be classified into sub-classes of meanings like 'perception,' 'recognition,' 'hearsay,' 'insist,' 'commit,' and so on. We observed cases in which certain sentence endings show correlation with certain tone types. For example, sentence endings indicating recognition of information (such as '-guna,' '-ji,' and so on) tend to have TL% instead of H% tones when the sentence is used as a request for response, while those indicating presumption (such as '-eulgeol/-eulgeol-yo') tend to have H% tones despite being declarative. We classified all of the sentence endings that can occur in conversational text (555 endings) into 44 semantic classes to construct a mapping table for modality tagging[3], which is partly shown in Table 2.

Table 2. Semantic class and sentence endings mapping sample.

Hearsay	-dae, -ndae, -neundae, -lae, -eulae, -jae, -ndae-yo, …
Commitment	-lge, -lge-yo, -eulge, -eulge-yo, …
Presumption	-lgeol, -lgeol-yo, -eulgeol, -eulgeol-yo, …
Reassertion	-nyanigga, -nyaniggayo, -danigga, -daniggayo, …
Negation	-llagu, -llaguyo, -ullagu, -ullaguyo, -gineun, …
Surprise	-daniyo, -danyo, -laniyo, -lanyo, -janyo, -janiyo, …

For sentence type tagging, we classified sentence types into 6 categories: statement, WH-question, yes/no-question, command, proposition, and exclamation. For speech act, we set 21 tags, such as 'give-information,' 'request-action,' 'commit,' and so on.

2. Experiment

A. Data

Using the tone type glossary of Table 1, the sentence-final tone types were manually labeled by listening and by visual inspection of the pitch contours using Praat [6], a speech analysis program. This was done by a single trained linguist in order to guarantee consistency. The only sentence with HLH% was deleted from this data set. Observing that LM% tones occur under the same conditions as L% tones which are typically used for statements, we assumed that the difference between L% and LM% tones is a difference in speaking style. It appears that LM% tones are used to express a speaker's confidence or certainty with a friendlier attitude than L% tones. But, given the rare occurrence of L% tones (1.7%) we merged them with the LM% tones. The linguistic features such as sentence type [7], speech act, and modality of sentences were tagged semi-automatically and manually corrected. The remaining data was randomly divided into a training set of 3,883 utterances and a test set of 970 utterances.

B. Method

For prediction modeling, a statistical method by probabilistic decision tree, or classification and regression tree (CART) was used. We performed an experiment using sentence type, speech act, and modality as major parameters in order to investigate which feature contributes most to the prediction of tone types. Other minor features such as type of punctuation, respect level, inclusion of WH-word, interjection, and so on were parameterized as well.

C. Results

The prediction accuracy results for the test set when using each of the linguistic features and their combinations are shown in Fig. 5. The results in Fig. 5 show that using the modality (MO) of sentence or its combination with other features as a parameter improves the accuracy up to 25% compared to cases in which other features or their combination are used (speech act (SA) and sentence type (ST)). The prediction accuracies for each tone type

3) In this work, only sentence endings were used for modality tagging. The other minor elements such as modal adverbs will be added in further research.

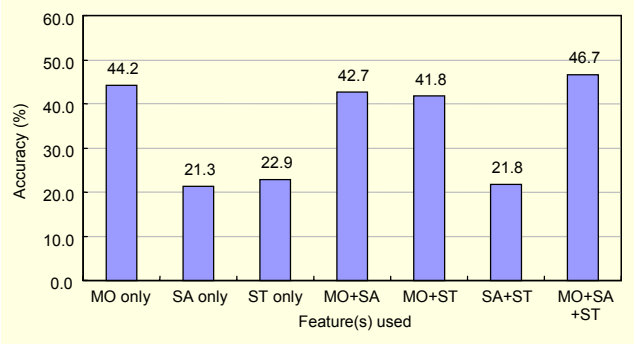

Fig. 5. Overall prediction accuracy using different features.

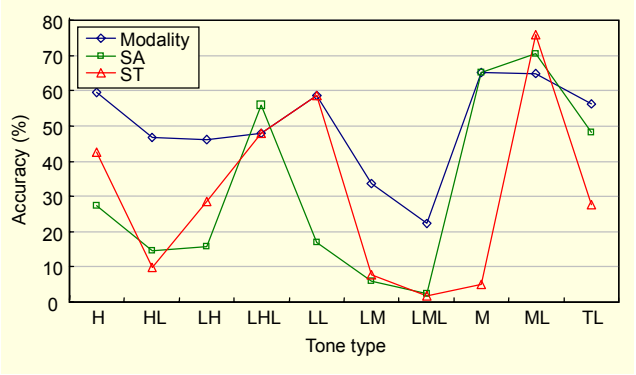

Fig. 6. Prediction performance using different features.

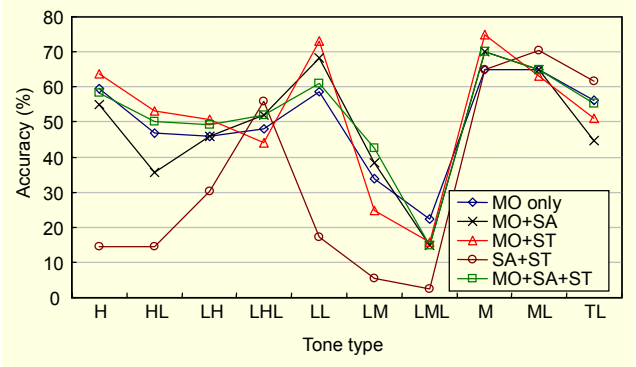

Fig. 7. Prediction performance using combined features.

using different features are shown in Fig. 6. It shows that the prediction accuracy improves for the H%, HL%, LH%, and LM% tones when the modality feature is used. Figure 7 shows that when combining modality with other feature(s), the prediction accuracy improves for the H%, HL%, LH%, LL%, and LM% tones.

In light of our current results, it is still too early to decide whether MO only or a combination of MO and SA, ST, or SA together with ST should be used as parameters since their relative performance does not significantly differ. We need to mention that the aim of this prediction model is not to perfectly produce the intonation types as labeled but to avoid producing unacceptable intonations. In these results, predicted tones which would have been acceptable as sentence-final tones were considered erroneous when different from their manual labels. If these are considered correct instead, the accuracy improves even more.[4] The low prediction performance of the LM% and LML% tones are thought to be a result of their interchangeability with each other.

IV. Conclusion

We proposed a list of sentence-final tone types for developing conversational-style Korean text-to-speech systems, based on their function and meaning. We have shown that modality has a closer correlation with tone types than sentence type or speech act and that prediction accuracy improves up to 25% when the feature of modality is used.

In our further research, we also need to introduce a modified performance metric which takes into account the interchangeability of tone types.

References

[1] S.A. Jun, *K-ToBI Labelling Conventions*, ver. 3.1, http://www.linguistics.ucla.edu/people/jun/ktobi/K-tobi.html, 2000.

[2] H.Y. Lee, "An Acoustic Phonetic Study of Korean Nuclear Tones," *Malsori,* The Korean Society of Phonetic Sciences and Speech Technology, vol. 38, 1999, pp. 25-39 (in Korean).

[3] F.R. Palmer, *Mood and Modality*, Cambridge University Press, Cambridge, 1986, p. 14.

[4] X. Huang, A. Acero, and H.W. Hon, *Spoken Language Processing,* Prentice Hall PTR, New Jersey, 2001, pp. 757-759.

[5] S.A. Jun and M.R. Oh, "A Prosodic Analysis of Three Sentence Types with 'WH' Words in Korean," *Proc. ICSLP'94,* Yokohama, Japan, 1994, pp. 323-326.

[6] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer," (Version 4.3.19) [Computer Program]. Retrieved July 20, 2005, http://www.praat.org/.

[7] C. Lee, G.G. Lee, and M.G. Jang, "Dependency Structure Applied to Language Modeling for Information Retrieval," *ETRI J.*, vol. 28, no. 3, June 2006, pp. 337-346.

4) When the accuracy was re-evaluated manually for the MO only case, considering all the acceptable tone types with 250 utterances in the test set, the accuracy improved from 44.2% up to 74%.