DISAMBIGUATION OF KOREAN UTTERANCES USING AUTOMATIC INTONATION RECOGNITION

*Tae-Yeoub Jang** *Minsuck Song*** *Kiyeong Lee****

*Centre for Speech Technology Research
University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, UK
**Department of English Language and Literature
***Department of Electronic Communication Engineering
Kwandong University, 522 Naigok-dong, Kangnung-shi, Kangwon-do, South Korea
email: tyjang@cstr.ed.ac.uk, {mssong,kylee}@kdccs.kwandong.ac.kr

ABSTRACT

The paper describes a research on a use of intonation for disambiguating utterance types of Korean[1] spoken sentences. Based on *tilt intonation theory* [8], two related but separate experiments were performed, both using the Hidden Markov Model training technique. In the first experiment, a system is established so that rough boundary positions of major intonation events are detected. Subsequently the significant parameters are extracted from the products of the first experiment, which are directly used to train the final models for utterance type disambiguation. Results show that the intonation contour can be used as a significant meaning distinguisher in an automatic speech recognition system of Korean as well as in a natural human communication system.

1. INTRODUCTION

Many experimental or theoretical linguistic researches on human languages so far have shown that the segmental structure is by no means sufficient in making clear the meaning of spoken utterances. [5] performed human perceptual tests and proved that syntactic disambiguation could be performed on the basis of prosody alone. [2] suggests some Korean phonological rules be re-described by using prosodic contexts for better explanation of phenomena.

In spite of such apparent consciousness of many linguists, phoneticians, and computer-speech scientists of this evident role of intonation, there has been little research which uses intonation to improve the quality of speech recognition system.

There are two main reasons. Firstly, a good method of extracting acoustic parameters for intonation recognition has not been discovered until recently. Secondly, variability of prosodic features have made it difficult to model the reliable intonation patterns. Compared with segmental structure of utterances, the range of variation of a specific prosodic feature is vastly stretched.

The *tilt intonation theory* [8] is an effort to overcome the first problem stated above. It tries to extract linguistically meaningful information from high-level knowledge of prosody and makes it easy to use. Once minor non-significant factors are removed, it extends the remaining important factors to represent various useful information sources in the way they can be easily put into practical experiments. A statistical analysis technique is a way of tackling the variability problem. As long as sufficient training tokens are provided, it becomes possible to have a relatively large range of information sources effectively compressed into a unit of a reasonably small size. That is the reason we use tilt parameters and a statistical recognition model, HMM, to model the intonation contour.

Given the background stated above on intonation, the starting point of our research is the belief that the major linguistic findings on intonation structure are also useful for practical speech recognition systems, if not as much as in human natural language. In many Korean sentences, two different meanings can frequently be contained in exactly the same single segmental structure, both orthographically and phonetically. In some cases a sentence can even contain a three-way ambiguity rather than two. Consider two Korean sentences for example.

1. **Two-way Ambiguity**
   nae-il-eun ppal-li toe-cyo[2]
   · Falling: ”It will be done faster tomorrow.”
   · Rising: ”Will it be done faster tomorrow?”
2. **Three-way Ambiguity**
   o-neul-eun eo-ti an-ka-ko cip-e iss-eoyo
   · Falling: ”I stay home today without going anywhere.”
   · Rising: ”Are you staying home without going anywhere?”
   · Level: ”Stay home and don’t go anywere!”

As illustrated, example 1 can be pronounced to emit the meaning of either statement or question, while example 2 can be realised in three ways: statement, question or request. But Korean hearers do not normally fail to recognise the speaker’s intended meaning, thanks to the sentence final intonational contour. So we assume that the utterance final intonational excursions of Korean has a considerable consistency and can be modelled statistically in a speaker-independent way.

[1] As intonational pattern of Korean varies dialect by dialect we restrict ”Korean” in this paper to ”Seoul Korean”, which is generally regarded as the standard dialect.

[2] In transcribing Korean pronunciation we follow the *Hangul Romanisation Standard* agreed between South and North Korean authorities in 1992.

This paper is composed of 4 main sections. The phonology of Korean intonation will be briefly introduced focusing on the shape of the intonation boundary which is directly relevant to our modelling. The method of collecting and arranging speech data will be reported. And then the procedure of two experiments and their results will be described and discussed in detail, followed by concluding remarks.

2. PHONOLOGY OF KOREAN INTONATIONAL STRUCTURE

There is general agreement among phonologists that an intonational phrase in Korean is marked by one major excursion of pitch contour and a boundary tone on the last syllable of the phrase. However, more than one minor intonation group can be located within one intonational phrase. Each minor phrase, frequently called an *accentual phrase* [2] [1], bears its own phrasal accent, usually close to its right-edge boundary tone. There are also possible pitch accents which take place non-uniformly depending upon many linguistic and para-linguistic factors including syntactic structure, speech rate, phonological focus, weight of the phrase, and pragmatic aspects of the individual speaker's intention.

Presuming the *Strict Layer Hypothesis* of prosodic constituents [6], an intonational phrase is composed of one or more accentual phrases. However, all the accentual phrases don't seem to be influential in determining the meaning of sentences. The role of all the accentual phrases but the last can rather be described as only preserving rhythmic structure of an utterance. For instance, a single sentence initial accentual phrase in a fast spoken utterance tends to be divided into two or more different accentual phrases when the same sentence is pronounced slowly. Consequently, the last accentual phrase or more specifically the last phrase accent of that accentual phrase along with its boundary tone is crucially responsible for the determination of meaning and disambiguation of an utterance. Moreover, the information of these two important elements is directly transmitted to the one-level higher intonational phrase finally making it work as a meaning distinguisher.

Therefore, we count important the information of the last accentual phrase, or equivalently the last pitch accent and the boundary tone of the intonational phrase. It doesn't mean in our research that all the other pitch accents are totally ignored. All of them are equally labelled and used as training tokens for the pitch accent detection. In other words, each of them helps in modelling pitch accent by providing values to be accumulated for statistical classification.

There is still no agreement on the number of meaningful intonation contour of Korean utterances. Phonologists or phoneticians suggest different numbers each other: 5 types in [2], 6 in [3], 7 in [4].[3] We used only three broad categories in our disambiguation experiments: *statement*, *question*, and *request*. The typical patterns of each type are illustrated as:

· *statement*: falling, pitch accent + falling
· *question*: rising, pitch accent + rising
· *request*: falling, neutral, neutral-falling

Though the above classification is certainly too general in linguistic terms considering a number of possible subdivisions for each class, we found that a great portion of Korean ambiguous colloquial sentences are related with those three types.

3. DATA COLLECTION AND VERIFICATION

14 speakers, all of whom are native speakers of Seoul Korean, participated in recording speech database. A script of 72 sentences was given to each subject with a brief instruction of recording. The reader knows in what utterance type the sentence should be pronounced through the punctuation marks of at the end of each sentence: '.' for statement, '?' for question, and '!' for request.

As all the data were recorded through each speaker's own equipment, some of them were first thought to be inappropriate to be used as training data. Nevertheless, none of them were excluded, in order to keep the quality of data as natural as they can be although, consequently, it might have caused the deterioration of the recognition accuracies reported later in this paper.

Forty three out of total seventy two sentences are ambiguous ones if they are represented only at the segmental level. Among them, 14 sentences are three-way ambiguous while 29 sentences are two-way ambiguous. If a sentence has a two-way semantic ambiguity it was recorded twice using a different intonation contour at each time. For example, when a sentence is ambiguous so that it can represent either **statement** or **question**, the reader was requested to read the sentence once as a question and once as a statement. Likewise, three-way ambiguous sentences are recorded three times, each time with a different tune.

Twenty nine segmentally non-ambiguous sentences were included in order to capture the general shape of intonation pattern for each utterance type regardless of whether the spoken utterance is ambiguous or not. That is, all non-ambiguous sentences are also used at the disambiguation experiment on section 5 to train each utterance type. In addition, this helped to get a little more natural speech data by decreasing the unnecessary tension or bias which could be shown by speakers when a fixed set of relevant data is provided.

After the data was digitised, it was verified by 3 native Korean speakers. There were a considerable number of mis-pronounced sentences as expected. Based on the verifiers' agreement the type annotation of 52 tokens was changed. But there are some sentences on which different judgements were made among the verifiers. We left these markings as they were first done and took a note of them for reference. 22 such cases might have degraded, if not to a great degree, the accuracy of utterance type recognition.

[3] The different number of classes among them dose not necessarily reflect their disagreement interpretation on an identical intonation pattern. Rather, their level of observation appears to slightly different one another. For instance, [2] mentions a little abstract phonological level while [4] takes into account more specific phonetic representation.

Language Model	Correctness(%)	Accuracy(%)
None	93.65	-67.03
Unigram	59.12	10.04
Bigram	68.12	15.10
FSN	70.16	38.01

Table 1: Event Recognition Accuracy

4. EVENT DETECTION

4.1. Experiment

In line with the tilt intonation theory two symbols **a** and **b**, called *intonation events*, are used in describing the intonation of spoken utterances. **a** stands for the pitch accent which can be paraphrased as an excursion of F0 located on a syllable within an intonational group.[4] The section between two events is expressed as **c** meaning *connection*, which doesn't contain any important information but plays a role of keeping continuity of event units. Combined symbols of **a** and **b** along with **r**ising or **f**alling notations were used to specify intonation contours such as **fb** (falling boundary), **rb** (rising boundary), **afb** (pitch accented falling boundary), and **arb** (pitch accented rising boundary).

To model each intonational events, three-state left-to-right continuous density Hidden Markov Models were constructed using HMM Tool Kit (HTK [10]). Speaker-normalised *F0* and *RMS energy* along with their *first and second derivatives* were calculated and saved in multi-dimensional vectors to define each model. Training was done by reading in all the relevant tokens, calculating rough estimates and then repeatedly re-estimating the means and variances of each vector until either convergence is reached or the designated iteration limit is reached. 304 utterances are used for training and each contains usually several pitch accents one boundary tones.

4.2. Result

Recognition was performed by running a Viterbi decoder over the 76 utterance tokens reserved only for tests. In estimating the accuracy of recognition, the duration of each recognised label is compared with that of the corresponding reference label in the reference file ready made by hand. When two labels overlap by 50%, recognition was marked is correct.[5]

The best result obtained so far is 70.16% correctness and 38.01% accuracy. The difference between correctness and accuracy lies in whether insertion errors were counted or not. In other words, a fairly large number of events which are absent from the hand-labelled file were found added at the auto-labelling stage. Some of those insertion errors could have been caused by inconsistency of hand labelling. There are many syllables for which it is hard to distinguish between a major accent location and an ignorable minor excursion.

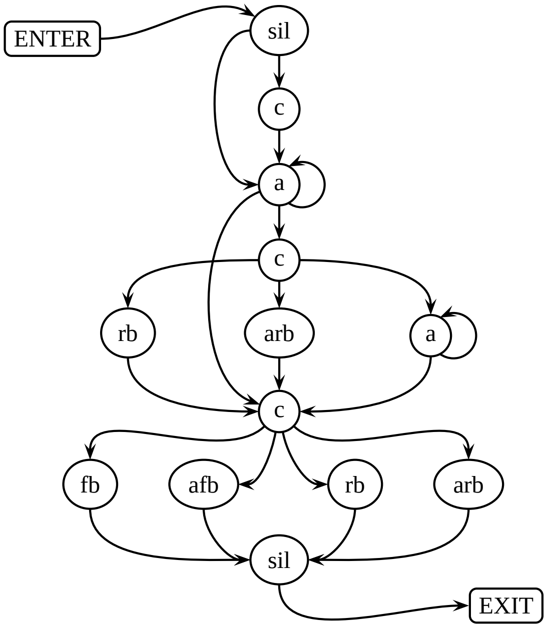

Figure 1: Finite State Grammar Model for Event Detection

In addition, the effectiveness of four different types of language model are compared. The results of recognition accuracy based on these models are shown in Table 1. When no language model is used and any sequence of events is allowed at any position, correctness shows the highest score. But as the corresponding negative accuracy indicates such high correctness is nothing but an unreliable result caused by a large number of inserted events all over the time span by the recogniser. Each reference event happened to overlap with one of such erroneous recognised tokens.

The artificially described finite state grammar model shows the best result in this research. The grammar, shown in Figure 1, was designed to reflect the phonologically verified intonational structure explained in 2, also on the basis of close investigation on the manually labelled intonational events. The most important part of Figure 1 is the description of the last few nodes, which forces the occurrence of a single boundary at the end of a sentence bearing the key information for utterance type disambiguation. While comparing boundary events with reference labels we found that the accuracy of two items **arb** and **rb** was as high as 89%. Considering the intonation pattern of Korean, such a result implies that most questions can be detected by event detection alone along with a simple finite state intonation grammar. If this information is used in a preliminary step before the main disambiguation we will get an improvement on the result shown later in this research.

5. DISAMBIGUATION

5.1. Tilt Analysis

The outputs of the event detection are the files with event labels and their positional values of starting and ending points. As stated

[4] Strictly, it is the case only in phonological terms that a pitch accent falls on a syllable, since acoustically it is quite difficult to detect the syllable especially when consonants are located between syllables.

[5] For the detailed description of the transcription comparison algorithm, see [7].

in [7], four parameters for each event are calculated in terms of amplitudes and durations of the F0 shapes, to extract linguistically meaningful information. They are *rise duration*, *fall duration*, *rise amplitude*, and *fall amplitude*. Rise duration is the distance between the starting point of the event and its peak, while fall duration is the distance between the peak and the ending point. Rise amplitude is the difference of F0 values between starting point and peak. Likewise, fall amplitude represents the difference of F0 values between start and ending point.

Though these parameters are acoustically meaningful in themselves, they are not optimal to be directly used as recognition parameters, since some information is represented somewhat redundant. Thus, more compact and efficient adjustment is performed in the form of *tilt* parameters which are briefly defined as follows:

· *Tilt Duration*: rise duration + fall duration
· *Tilt Amplitude*: |rise amplitude| + |fall amplitude|
· *Tilt Itself*: overall shape of the event

The last item is an abstract representation whose value is generated by both F0 amplitude and duration.

5.2. Utterance Type Modelling

After automatically demarcating event boundaries of provided training data by running the event detector, tilt parameters are calculated as explained above. All the values are combined in a single file and used as training inputs.

Using the parameters extracted as explained in the previous section, three three-state Hidden Markov Models are trained through the standard Baum-Welch algorithm, each representing an utterance type. 449 utterances are inputted as training tokens along with corresponding type annotations and the other 251 tokens are used for testing. None of the test tokens have been used before for training either event detector or utterance type recogniser although the speakers of each data set may overlap. Event detection, parameter calculation, and utterance type recognition are performed automatically over test each test token and the final judgement is made.

5.3. Results and Discussion

The best overall result obtained through the utterance type recognition is 56.97%. For each utterance type, 50.5% statement, 70.93% question, and 50% request sentences are correctly detected. When we perform the estimation only for the ambiguous sentences we get a little better result. 68.04% of the utterances are correctly disambiguated. The improvement seems reasonable considering that most ambiguous sentences, whether two-way or three-way, contains interrogative meaning. That is, the ambiguity only between the other two utterance types (statement and request) are relatively rare and the difficulty of the disambiguation in this case is naturally avoided.

Better results are expected in the future. Data extension is necessary in both quantity and quality. More natural and practical data such as spontaneous dialogue or telephone conversation need to be collected in a greater amount for better comparison and enhancement of the recognition performance. Voices of more people need to be collected as a train data in order to construct a more reliable speaker independent model.

A larger database doesn't necessarily mean a subsequent deterioration in accuracy. For example, when the experiment is performed over a database composed of natural dialogues, there is at least an additional useful cue to help detect utterance type. It is the nature of dialogue that the type of each utterance may well be closely related with neighbouring utterances. This information can be used as in a probabilistic N-gram language model while recogniser is being run. The effect of this method has already been verified successfully in [9].

For further tests of the usefulness of utterance type recognition the number of types needs to be extended. Phonological analyses, as described in section 2, suggest a more refined classification than the three types dealt with here. Those various high-level classifications need to be tested as there has been no serious study yet on the best number of utterance types practically applicable to the speech recognition system.

6. CONCLUSION

Through the two experiments we verified two main hypotheses. First of all, we saw that utterance final intonational information can be used as a meaning distinguisher of a speech recognition system of Korean. We also confirmed that tilt intonation parameters are useful in capturing linguistically meaningful prosodic features. The HMM training was useful for modelling highly variable prosodic features as well.

7. REFERENCES

1. Mary E. Beckman and Janet B. Pierrehumbert. Intonational structure in Japanese and English. *Phonology Yearbook 3*, pages 255–309, 1986.
2. Sun-Ah Jun. *The phonetics and phonology of Korean prosody*. PhD thesis, The Ohio State University, 1993.
3. Hee San Koo. *An Experimental Acoustic Study of the Phonetics of Intonation in Standard Korean*. PhD thesis, University of Texas at Austin, 1986.
4. H. Lee. *Korean Phonetics*. Taehaksa, Seoul, Korea, 1996. (in Korean).
5. P. J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. The use of prosody in syntactic disambiguation. *Jounal of the Acoustical Society of America*, 90(6):2956–2970, 1991.
6. E. Selkirk. *Phonology and Syntax: The Relation between Sound and Structure*. MIT Press, Cambridge, Massachusettes, 1984.
7. Paul Taylor. Analysis and synthesis of intonation using the tilt model. Draft, 1997.
8. Paul Taylor and Alan W. Black. Synthesizing conversational intonation from a linguistically rich input. In *2nd ESCA/IEEE Workshop on Speech Synthesis*, 1994.
9. Paul Taylor, Hiroshi Shimodaira, Stephen Isard, Simon King, and Jaqueline Kowtko. Using prosodic information to constrain language models for spoken dialogue. In *International Conference on Spoken Language Processing96*, Rhode, Greece, 1996.
10. S. Young, J. Jansen, J. Ollason, and P. Woodland. *HTK Book*. Entropic, 1996.