Automatic Rule Generation for Linguistic Features Analysis Using Inductive Learning Technique
—Linguistic Features Analysis in TOS Drive TTS System—

Shigenobu Seto, Masahiro Morita, Takehiko Kagoshima, and Masami Akamine

TOSHIBA Corporation, Kansai Research Laboratories
6-26, Motoyama Minami-cho, 8-chome, Higashinada-ku, Kobe, 658-0015, JAPAN

ABSTRACT

The linguistic features analysis for input text plays an important role in achieving natural prosodic control in text-to-speech (TTS) systems. In a conventional scheme, experts refine suspicious if-then rules and change the tree structure manually to obtain correct analysis results when input texts that have been analyzed incorrectly. However, altering the tree structure drastically is difficult since attention is often paid only to the suspicious if-then rules. If earlier rule-tree structure is inappropriate, any attempt to improve the performance may be limited by the stiffness of the structure.

To cope with these problems, the new development scheme generates analysis rules by using C4.5 [1], where an if-then rule-tree structure is generated by off-line training. The scheme has the advantage that since the generated rule-tree structure is simple, the rules are easier to maintain. The scheme is applied to generating four types of analysis rule-trees: rules for forming accent phrases, rules for determining accent position, rules for analyzing syntactic structure, and rules for pause insertion. An experimental evaluation was performed on these four rules. The accuracy was 96.5 percent for the accent phrase formation, 95.5 percent for the accent positioning, 87.0 percent for the pause insertion, and 88.3 percent for the syntactic analysis despite using small training data. These results indicate the validity of the scheme. The new scheme is used for developing linguistic features analysis rules in a Japanese TTS system, TOS Drive TTS [3].

1. INTRODUCTION

The linguistic analysis and the prosodic pattern control are significant components in generating natural synthesized speech from input text. The linguistic analysis for input text extracts linguistic features, such as accent position, syntactic structure, and so on, which are essential factors for determining the shape and the height of pitch pattern. For example, the pitch pattern where the pitch frequency falls around n-th mora is selected if the accent is located on the n-th mora.

Achieving accurate linguistic features analysis for a wide variety of texts requires carefully designed analysis rules and iterative refinement of the rules because of the following two reasons. First, accent position of a word often shifts when the word concatenates to other words. The list of linguist made rules which shows how the accent positions shifts is not complete. It includes many exceptions and is ambiguous about priority among the rules [4]. Second, there are sometimes several possibilities of acceptable accent position and syntactic structure. Word accents often vary from generation to generation. More than two possibilities of accent position can be acceptable, or the acceptable possibilities may differ between generations. To be able to follow the transition of word accent characteristics, the accent rules should be easy to reconstruct.

In the conventional scheme of developing rules, the list of rules is implemented, after that experts manually refine the if-then rules and their structure. There are several problems in that scheme. First, drastically altering the rule-tree structure is difficult if the structure of the list of rules is inappropriate. Second, iterative improvement of the rules is a labor-intensive process, since the result of the rule changes is hard to evaluate before the changes are made. Last, the conventional scheme is apt to lack flexibility and scalability for different applications. A specific application of the TTS system may become clogged with unnecessary analysis rules if the TTS system uses only limited grammatical structure.

To overcome these problems, the new scheme for generating rules is employed using one of the inductive learning techniques, C4.5 [1]. It generates a decision tree (rules) from training data set like CART [5][6][7]. But, the C4.5 generates a simpler rule tree since nodes of the decision trees can have more than two branches while the tree that is generated by CART is a binary tree. Four analysis rules are automatically generated using the C4.5, to be used for online linguistic features analysis: rules for forming accent phrase from morphological analysis result, rules for determines accent position, rules for obtaining syntactic structure information, and rules for determining pause insertion position.

An experimental evaluation was performed on the four rules by counting number of accent phrases, whose feature was estimated correctly against the original phrases. The accuracy was 96.5 percent for the accent phrase formation, 87.2 percent for the accent positioning, 87.0 percent for the pause insertion, and 88.3 percent for the syntactic analysis. These results show the validity of the scheme, and the performance could be improved by pre-processing or post-processing using a few heuristic rules.

2. GENERATING RULES USING C4.5

The C4.5 [1] is an inductive learning technique. It generates classification rules from training data that consist of pairs of an input attribute vector and an output appropriate class. The obtained classification rules assign new attribute vectors to the classes. It is a powerful technique not only for learning tasks that have discrete attributes, but also for tasks with continuous attributes [2].

In the training process, the C4.5 first generates an initial decision tree. Each node has a test question to an attribute value of the input vector, and each leaf indicates the class to which the vector should be assigned. The fact that the C4.5 allows more than two divisions for a node makes the tree structure simpler and helps the system developers understand the obtained rules more easily. After generating the initial decision tree, the C4.5 prunes branches away in order to avoid over-fitting to training data and to make the rules more robust against open data. It can also generate a set of production rules from the simplified rule-tree.

3. AUTOMATICALLY GENERATING LINGUISTIC ANALYSIS RULES

The C4.5 is employed for constructing the following four types of analysis rules in the TOS Drive TTS system.

- Accent phrase boundary determination (Accent phrase formation) rule
- Mora-by-mora tone prediction (Accent type determination) rule
- Modified word distance prediction rule (Syntactic structure analysis)
- Pause insertion rule

Figure 1 shows a block diagram of the linguistic analysis in the TTS system. The morphological analysis converts input text to a sequence of morphemes, which is followed by the linguistic features analysis. The linguistic features analysis has four stages: accent phrase formation, accent type determination, syntactic structure analysis, and pause insertion. In the accent phrase formation, a sequence of morphemes is broken into accent phrases. An accent phrase is a basic prosodic unit that has one accent within it. The accent phrase formation rule is used for determining whether each morpheme boundary is an accent phrase boundary. In the accent type determination, the accent position for each accent phrase is selected from the possible candidates. A new method is applied for determining the accent position, as will be mentioned later. The accent position candidates of an n-morae accent phrase in Japanese Tokyo dialect are type 0 (unaccented), type 1 (the accent is located on the first mora) and so on to type n−1 (the accent is locates on the n-1 th mora). The mora-by-mora tone prediction rule is used for estimating the scores of each candidate, and the best-scored candidate is selected. In the syntactic structure analysis, *modified word distance* is predicted. When the *i*-th accent phrase AP_i modifies the *j*-th accent phrase AP_j, the modified word distance of AP_i is defined as j−i. The modified word distance prediction rule is used for predicting the value. In the pause insertion, it is determined if a pause should be inserted on each accent phrase boundary and how long the pause should be, using the pause insertion rules. The results of the four-stage processing are tagged to the accent phrases and are passed to the prosodic control component for determining the prosodic pattern.

After the morphological analysis each morpheme is tagged with morpheme attributes, such as pronounciation, part of speech, the number of mora, accent position of the morpheme, conjugation, type of character, accent shift attributes, information for forming compounds and so on. Among these attributes, a set of attributes are selected and used as input vector for generating the rule tree. In order to obtain an appropriate rule tree and to avoid over-fitting, each attribute is checked whether it is effective. If an attribute is found to be located too far from the root node and the attribute is used for only trivial rules, the attribute is removed from the input vector.

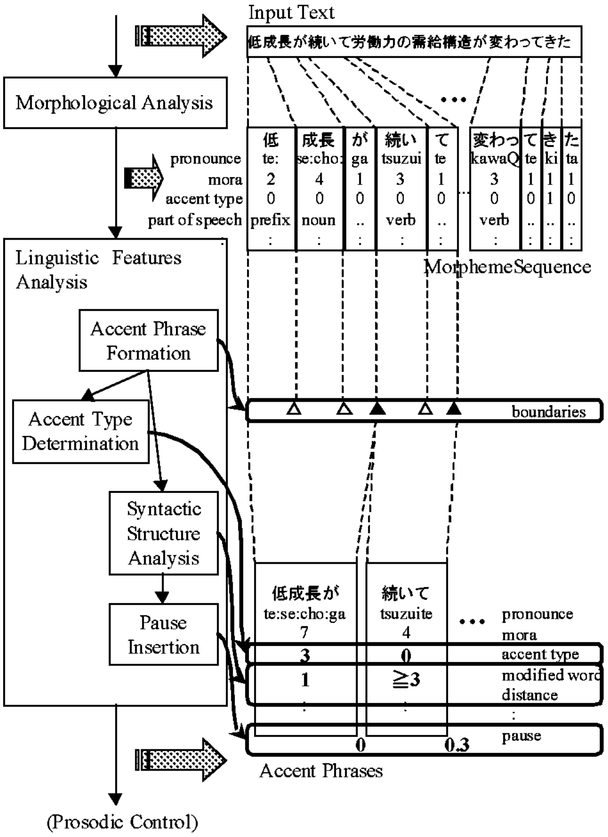
Input Text
低成長が続いて労働力の需給構造が変わってきた
Morphological Analysis
低 成長 が 続い て 変わっ て き た
pronounce te: se:cho: ga tsuzui te kawaQ te ki ta
mora 2 4 1 3 1 3 1 1 1
accent type 0 0 0 0 0 0 0 1 0
part of speech prefix noun .. verb .. verb .. .. ..
MorphemeSequence
Linguistic Features Analysis
Accent Phrase Formation
Accent Type Determination
Syntactic Structure Analysis
Pause Insertion
boundaries
低成長が te:se:cho:ga 7
続いて tsuzuite 4
pronounce
mora
accent type
3 0
modified word distance
1 ≧3
pause
0 0.3
Accent Phrases
(Prosodic Control)

Figure 1: Block diagram of linguistic analysis in TOS Drive TTS system.

3.1 Accent Phrase Formation

To form an accent phrase, the morpheme attributes from rules made by linguists [4] are used as input, such as part of speech, the number of mora, accent shift attributes, accent position of the morpheme, conjugation, information for forming compounds, and so on. The morpheme attributes of four morphemes (two before and two after the boundary) are used for determining whether each morpheme boundary is an accent phrase boundary.

For the accent phrase formation rules, 8539 morpheme boundary data (about 1250 sentences) were used for training and evaluation. The obtained rules performed at 96.5 percent accuracy. The rate is cross-validated (successive training on 90 percent of the data and testing on 10 percent).

3.2. Mora-by-Mora Tone Prediction

In the accent type determination, a new method based on mora-by-mora tone prediction is exploited for determining the accent position. In the conventional method, accent position of the accent phrase is determined with word-by-word rules, where morpheme attributes such as accent position of morpheme and part of speech are used for determination. However, the rules have many exceptions. For example, shift of accent position often occurs around mora phonemes (N, Q, :). Priority between the rules and the exceptions are ambiguous [4].

The new method consists of two stages. In the first stage, H (high) and L (low) tone estimation for each mora in accent phrase is performed where the probability that the mora is uttered at high/low tone (p_H / p_L) is calculated. The C4.5 is employed for generating rules for estimating the probability. Figure 2 shows an example of probability of tone H for each mora of an accent phrase, "走れなかったので (ha shi re na ka Q ta no de)".

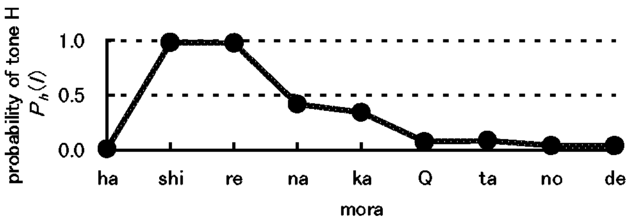

Figure 2: Example of predicted tone H score for accent phrase ("走れなかったので" : ha shi re na ka Q ta node).

In the second stage, scores for each candidate of accent type (accent position) are calculated. For example, in the case of a four-mora accent phrase, the scores for type-0, type-1, type-2, type-3 are defined as

$$S(0)= p_L(1)\, p_H(2)\, p_H(3)\, p_H(4),$$

$$S(1)= p_H(1)\, p_L(2)\, p_L(3)\, p_L(4),$$

$$S(2)= p_L(1)\, p_H(2)\, p_L(3)\, p_L(4),$$

$$S(3)= p_L(1)\, p_H(2)\, p_H(3)\, p_L(4),$$

respectively. Then the accent type that has the maximum score is selected. In the case of the example of Figure 2, type-3 is determined as the accent type of the accent phrase.

For the mora-by-mora tone prediction rules, the morpheme attributes and the position where the mora is located within the morpheme and the accent phrase and type of phoneme are used as input attributes.

For the mora-by-mora tone prediction rules, 43156 tone data of morae (about 1250 sentences and 1800 phrases, which include 10419 accent phrases) were used for training. The obtained rules were evaluated by 765 accent phrases (100 sentences open data). The determination accuracy was 96.5 percent by counting the numbers of correctly determined accent phrases.

4.2 Modified Word Distance

To utilize syntactic structure information in pause placement and pitch pattern control, predicting *modified word distance* is performed. When i-th accent phrase AP_i modifies j-th accent phrase AP_j, the modified word distance is defined as j–i. Figure 3 shows examples of modified word distance. In the currently implemented version, a modified distance has an upper limit of 3, because it makes little difference if the distances are long.

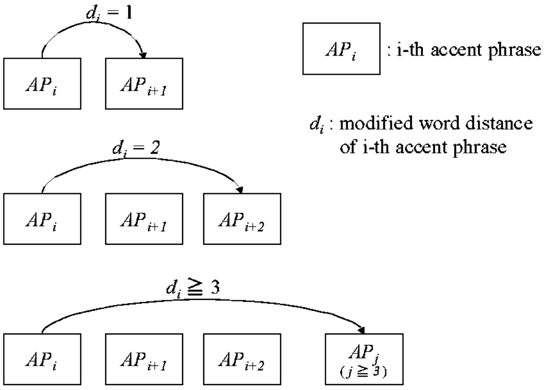

Figure 3: Modified word distance of accent phrase.

For the modified word distance prediction rule, attributes of the first and the last morpheme of the accent phrases, AP_i, AP_{i+1}, AP_{i+2}, AP_j are used. The numbers of morae of these accent phrases are also used for the input attributes.

For the modified word distance prediction rules, 5853 pairs of accent phrases (983 sentences) were used for training and evaluation. The obtained rules performed at 88.3 percent accuracy (The rate is cross-validated).

4.3 Pause Insertion

For determining pause insertion, the attributes of the morpheme before the boundary, the modified word distance of the accent phrases before/after the boundary, the number of morae of the accent phrase before/after the boundary, and the type of symbol on the boundary are used as input attributes. Figure 4 shows an example of the obtained rule-tree.

For the pause insertion rules, 4334 data of accent phrases boundary (693 sentences) were used for training and evaluation. The obtained rules performed at 88.3 percent accuracy (The rate is cross-validated).

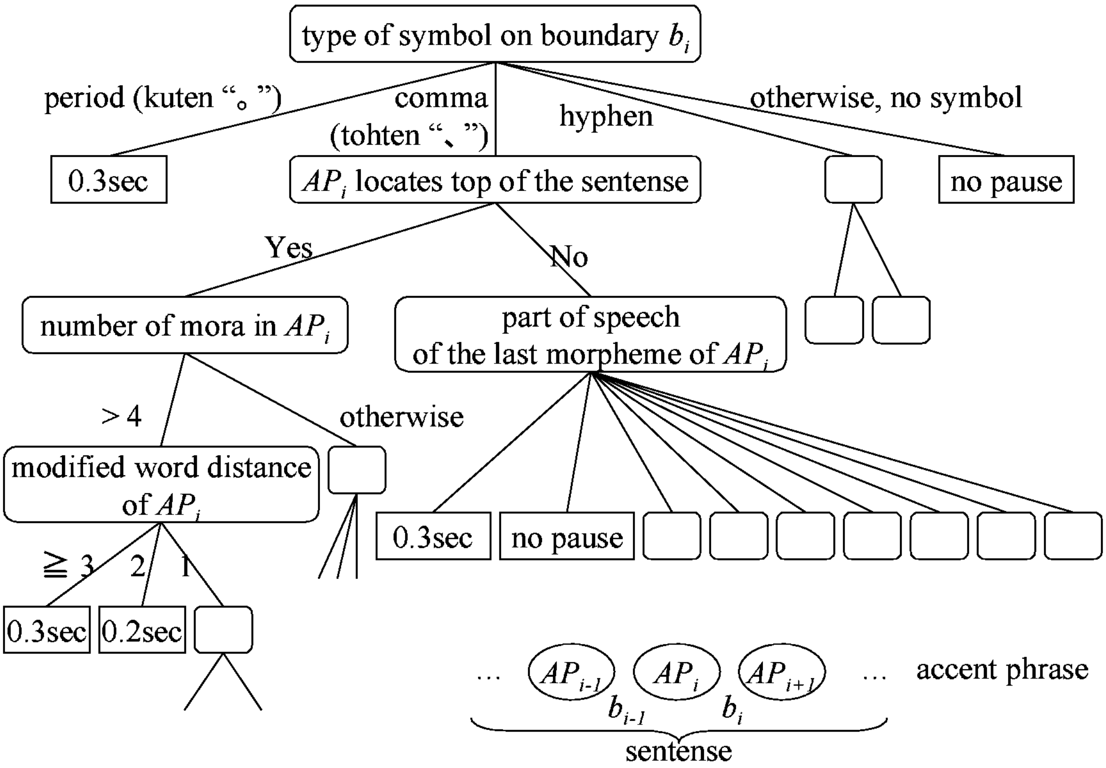
type of symbol on boundary b_i
period (kuten "。")
comma (tohten "、")
hyphen
otherwise, no symbol
0.3sec
AP_i locates top of the sentense
no pause
Yes
No
number of mora in AP_i
part of speech of the last morpheme of AP_i
> 4
otherwise
modified word distance of AP_i
≧ 3
2
1
0.3sec
no pause
0.3sec
0.2sec
… AP_{i-1} AP_i AP_{i+1} … accent phrase
b_{i-1}
b_i
sentense

Figure 4: Example of pause insertion rule (partially displayed).

4. CONCLUSION

The new scheme for developing linguistic features analysis rules using the C4.5 is described. It reduces time for reconstructing rules. The generated rule tree structure indicates priority of each rule. Rules that are closer to the root are more important than the rules further down (closer to the leaves). The generated rule tree has simple structure, which brings easiness for maintenance of the rules to system developers. The results of experiments show the validity of the scheme even though the rules are generated from a small training data set. The performance could be improved by pre-processing or post-processing manually using a few heuristic rules.

5. REFERENCES

1. Quinlan, J. R. *C4.5 : Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.
2. Quinlan, J. R. "Improved Use of Continuous Attributes in C4.5," *Journal of Artificial Intelligence Research*, **4**, pp. 77–90, 1996.
3. Kagoshima, T., Motita, M., Seto, S., Akamine, M. "An F_0 Contour Control Model for Totally Speaker Driven Text to Speech System," *Proceedings ICSLP 98*, Sydney, 1998.
4. Akinaga, K. "共通語のアクセント (Kyoutsuugo no Accent)," 日本語発音アクセント辞典 (Nihongo hatsuon accent jiten), pp. 45–90, 1966. (in Japanese)
5. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and Regression Trees*, Belmont, CA, Wadsworth, 1984.
6. Sproat, R., Hirschberg, J., and Yarowsky, D. "A Corpus-Based Synthesizer," *Proceedings ICSLP 92*, pp. 563–566, Banff, 1992.
7. Hirschberg, J. "Pitch Accent in Context: Predicting Intonational Prominence From Text," *Artificial Intelligence*, **63**, pp. 305–340, 1993.
8. Andersen, O., Kuhn, R., Lazaridès, A, Dalsgaard, P., Haas, J., Nöth "Comparison of Two Tree-Structured Approaches for Grapheme-to-Phoneme Conversion," *Proceedings ICSLP 96*, pp. 1700–1703, Philadelphia, 1996.