SEGMENT SELECTION METHOD BASED ON TONAL VALIDITY EVALUATION USING MACHINE LEARNING FOR CONCATENATIVE SPEECH SYNTHESIS

Akihiro Yoshida, Hideyuki Mizuno, and Kazunori Mano

NTT Cyber Space Laboratories, NTT Corporation, 1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa, 239-0847 Japan

ABSTRACT

This paper proposes a speech segment selection method based on machine learning for concatenative speech synthesis systems. The proposed method has two novel features. One is its use of Support Vector Machine (SVM) to estimate the subjective correctness of pitch accent with respect to each accent phrase of possible candidate speech segments. The other is its use of a determination function to identify the best segment based on SVM output. The determination function involves two assessments; one counts the number of each sign of SVM output and the other compares the distance values. The sign of SVM output is generally used to classify target objects, but the distance SVM output also represents important information. An experiment that assesses SVM performance for Japanese accent validity shows that its accuracy is 81%. To confirm the effectiveness of the proposed segment selection method, preference tests are conducted. The test indicates that the proposed method can yield Japanese synthesized speech with more natural intonation than the conventional method that uses only target cost and concatenation cost.

Index Terms— concatenative speech synthesis, segment selection, accent, machine learning

1. INTRODUCTION

TTS systems based on the concatenative speech synthesis framework can create very natural speech without prosodic modification if a huge speech corpus is available[1]. The speech segments are selected by a cost function designed so that the concatenative synthesis systems output high quality synthesized speech[2][3][4][5]. Recent cost calculation functions offer synthesized speech quality similar to human natural speech and the inverse correlation between such costs and speech quality has been confirmed[6][7].

Conventional cost calculation functions consist of target features and continuity features; the sub-cost relevant to prosodic features is calculated from the similarity between target prosodic features and those of speech segments' within the local prosodic range such as a phoneme or syllable. Selecting a combination of synthesis segments to minimize the sum of the costs can synthesize speech with small distortion with regard to target prosody. However, even if the selected speech segments match the target in the local range of phonemes, it isn't assured that the synthesized speech created by these speech segments is high quality. For example, if the tonal fluctuation of the Japanese synthesized speech yields several auditory accent falls, the quality of the synthesized speech is unacceptable because Japanese phrases have only one accent fall at most. In fact, the quality of Japanese synthesized speech created by combining speech segments with minimum total cost is sometimes inferior to the quality of those created with other speech segments yielding higher total cost. A preliminary subjective experiment on speech synthesized by the conventional method was carried out to determine which auditory factors provide the listener with a negative impression. It showed that among articulation, pitch accent, duration, and power, pitch accent is the most significant factor.

For this reason, the first priority is to propose a new selection method that can output synthesized speech with the right accents. The selection method should provide the best speech segments while considering prosodic features within an extensive range such as phrase or sentence.

In Section 2, an overview of the TTS system with the proposed speech segment selection method is described. Section 3 details the machine learning of the Support Vector Machine (SVM) for the tonal evaluation validity and its accuracy is shown. Section 4 describes the proposed segment selection method; the results of a subjective experiment confirm the advantages of the proposed method. Subjective evaluation results are discussed in Section 5.

2. OVERVIEW OF TTS SYSTEM BASED ON PROPOSED METHOD

In concatenative speech synthesis systems, transforming the input text into synthesized speech involves text analysis, target prosody generation, and speech segment selection. Our solution is a new speech segment selection method that introduces an evaluation scheme targeting tonal validity; it employs SVM to identify the most appropriate combination of speech segments. The proposed method consists of three steps as follows.

In the first step, N-best sequences of speech segments are identified by conventional segment selection based on total cost, because total cost is strongly and inversely correlated with synthesized speech quality, and higher candidates of sequences of speech segments for the cost are more likely to yield good combinations of speech segments.

The second step is to use SVM to estimate the subjective tonal validity of the N-best sequences of speech segments from their global F_0 features.

The last step is to select the optimal sequence of speech segments from all sequences based on the tonal validity estimated by SVM.

A flowchart of the proposed segment selection process is shown together with that of the conventional segment selection process in Figure 1. It shows the difference between the proposed segment selection method and the conventional method. The dashed line delineates the processes added by the proposed method.

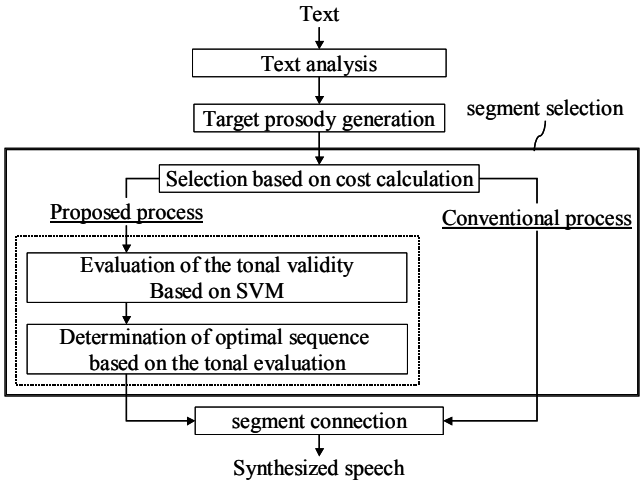

Figure 1. Flowcharts of concatenative speech synthesis systems.

3. SVM FOR EVALUATION OF TONAL VALIDITY

In this section, we discuss how to evaluate the tonal validity of synthesized speech. We employ SVM[8] as the tool to automatically evaluate the tonal validity, since it has high decision performance and can estimate nonlinear boundaries dividing target objects. We trained SVM with polynomial kernel of degree 3.

3.1. SVM for tonal validity evaluation of each accent phrase

We designed SVM features to evaluate the tonal validity of each accent phrase. Accent phrase is the fundamental unit of prosody in Japanese. The Japanese language expresses accent by the high-low change of the F_0 frequencies of adjacent mora and accent type is determined by the combination of words that compose an accent phrase. Therefore, in Japanese concatenative speech synthesis, synthesizing speech with appropriate intonation equals selecting the sequence of speech segments that yields the correct accent.

To check whether the accent of speech is correct or not, it is necessary to confirm that the F_0 frequency properly decreases at the right mora position. However, accent is, after all, just an auditory phenomenon and just a quantitative assessment of F_0 contour is not enough to identify the tonal pattern that humans perceive as accent. Since SVM offers non-linear classification, it is able to automatically discriminate the validity of accent in the same way as humans.

To evaluate whether the accent of an entire accent phrase, which is a sequence of speech segments, is correct or not, human auditory decisions are given as the target value and several F_0 features and qualitative values such as accent type are used as the SVM features as shown in Table 1. Each accent phrase takes its value from the SVM feature. However, as the SVM feature, we set up the maximum and the minimum value with respect to the gap of F_0 and the difference of center F_0 for three sections. The first section is the beginning part of accent phrase in which F_0 rises.

Table 1. SVM features

Kind of feature	Feature of SVM
qualitative (discrete-valued)	- existence of pause at anteroposterior position - accent type of anteroposterior accent phrases - accent type and mora count of accent phrase
quantitative (continuous-valued)	- regression coefficient of F_0 within mora - the gap of F_0 at mora boundary - difference of center F_0 between adjoining mora

Table 2. Data amount corresponding to accent phrase

	Positive data	Negative data	total
Learning data	4850	6724	11574
Test data	2901	1056	3957

Table 3. Auto-detection of accent validity

		Correct decision by human	
		Positive	negative
System decision	Positive	A : 2476	B : 328
	negative	C : 425	D : 728

The second section, the steady F_0 part, extends to the accent hole. The third section is the falling F_0 part after the accent hole. We expect to detect abnormal F_0 transitions by checking the two extreme values.

3.2. Experiment confirming SVM performance

As learning data and test data for SVM, synthesized speech samples were prepared. Each accent phrase was rated by an expert accustomed to evaluating accent as either true or false in terms of (natural) accent. Since a concatenative TTS system outputs a significantly larger amount of synthesized speech with true accent than with false, it is easy to collect positive data, which is synthesized speech with correct accent, while it is difficult to collect sufficient amounts of negative data, which is synthesized speech with incorrect accent, as learning data. The SVM constructed from such learning data couldn't achieve highly accurate classification. Consequently, to prepare a large quantity of false data, we used synthesized speech made by two kinds of concatenative TTS systems that had a small speech database and whose cost calculation function was deliberately tuned improperly. Test data consisted of synthesized speech output by a proper concatenative TTS system. The quantities of learning data and test data are described in Table 2.

We used these learning data to develop our SVM and conducted an experiment to evaluate the validity of accent of the resulting synthesized speech. The result is shown in Table 3. Positive means that the test data is classified as speech with correct accent and negative means that it is classified as speech with incorrect accent. The recall ratio of true data is 85% [A / (A+C)]. This indicates that SVM can correctly identify fair accents. On the other hand, the recall ratio of false data is 69% [D / (B+D)]. Its accuracy, i.e. the rate at which SVM can correctly judge the validity of accent, is 81% [(A+D) / (A+B+C+D)].

4. SEGMENT SELECTION USING DETERMINATION FUNCTION

4.1. Outline of segment selection

The proposed process of segment selection shown in Figure1 outputs the best speech segment sequence, i.e. that which has minimum total cost among the sequences judged to have acceptable accent by a determination function.

At first, the top 30 sequences of speech segments, as N-best, ranked in terms of total cost are selected. This was done because a preliminary experiment showed that the sequences outside the top 30 sequences ranked by total cost rarely yield high quality synthesized speech. Next, SVM evaluates each accent phrase in each sequence of the speech segments and tags each with either positive or negative sign and a distance value in terms of accent correctness. Finally, the determination function decides the optimal sequence of speech segments.

This paper examines two determination functions of Functions A and B. Function A identifies the optimal sequence of speech segments on the basis of the sign, positive or negative, of SVM output. Function B decides the optimal sequence of speech segments on the basis of the distance value of SVM output.

4.2. Determination functions

4.2.1. Determination Function A

Function A identifies the optimal sequence of speech segments on the basis of the sign, positive or negative, of SVM output. The sequence with the maximum number of positive accent phrases is selected. If some sequences have the same maximum number, the one with the minimum total cost is selected.

4.2.2. *Determination Function B*

Function B decides the optimal sequence of speech segments on the basis of the distance value of SVM output. This is expected to yield more accurate judgments than that possible when using only signs. The sequence with the maximum number of accent improvements is selected. The number of accent improvements is defined as the number of accent phrases in a sequence whose SVM distance value is sufficiently larger than that of the corresponding accent phrase in the sequence with minimum total cost.

Figure 2 shows the precision for positive test data and negative test one on the decision boundary of the SVM. In general, the decision boundary of SVM is the zero distance value. This figure shows the change of precision when shifting the decision boundary. The horizontal axis indicates the decision boundary and the vertical axis plots the precision of SVM for the positive test data and negative test data decisions. The distance values over or under threshold are limited to threshold values respectively, because too high or low distance values are not a reliable indicator of accent correctness as shown in Figure 2. Both the precision of the positive and of the negative test data saturate (more than 1.3 and less than -1.1), so we set the upper threshold to 1.3 and the lower threshold to -1.1.

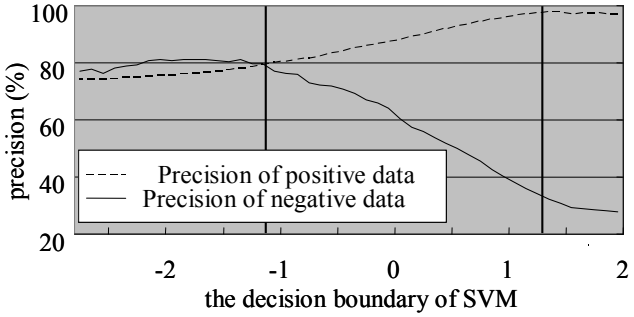

Figure 2. Precision of decision function thresholds

Function input was 30 speech segment sequences. The sequences were arranged according to the total cost. The scheme of Function B is shown in Figure 3. For the distance values $V(i,j)$, suffix i means the order of a sequence in the 30 sequences and suffix j means the accent phrase number in the sequence. If the function judges that the tonal validity is improved, the score $S(i,j)$ of the accent phrase is set at a positive value. If the function judges the quality to be degraded, the score $S(i,j)$ of the accent phrase is set at a negative value. In the other case, the score $S(i,j)$ is set at zero. The following is the algorithm of Function B.

(1) At first step, score $S(i,j)$ is calculated as follows.
If $V(i,j) - V(1,j) > Th_{u1}$, then $S(i,j) = 1$,
else if $V(i,j) - V(1,j) > Th_{u2}$ and $V(i,j) = 1.3$,
then $S(i,j) = 1$,
else if $V(i,j) - V(1,j) < Th_l$, then $S(i,j) = -1$,
else $S(i,j) = 0$.

(2) Next step, total score, $TS(i)$, is calculated as the summation of $S(i,j)$ for all accent phrases in i-th order sequences.

(3) Last, the sequences, whose order is lowest in sequences with the highest $TS(i)$, are output.

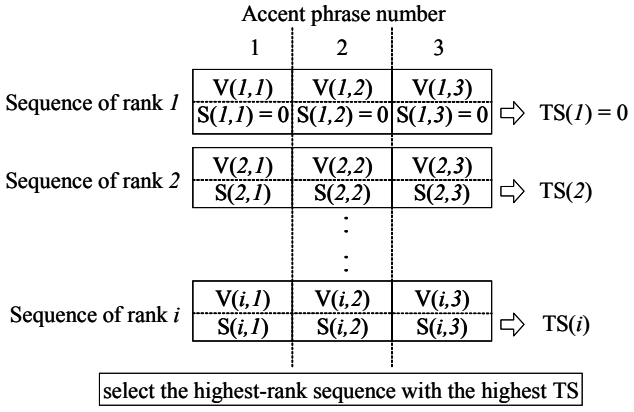

Figure 3. Scheme of Function B for the case that a sequence consists of three accent phrases

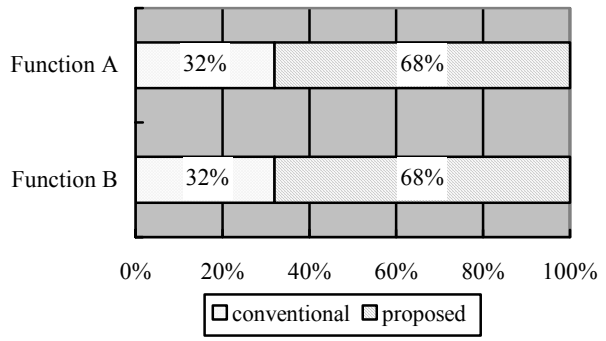

Figure 4. Results of preference test

5. EVALUATION OF PROPOSED METHOD

5.1. Pair comparison test

A preference test was carried out to confirm that the proposed method can output fair accent synthesized speech. As the stimuli for the test, three kinds of speech synthesized by the proposed selection methods, based on either Function A or B, and by a conventional selection method wherein Th_{u1} is set to 1.3, Th_{u2} is set to 0.5 and Th_l is set to 0.5. All conditions were identical except for segment selection. Open texts of news stories, each of which consisted of three accent phrases, were prepared for the evaluation.

If the proposed method selects the lowest total cost sequence as the optimal sequence, the output of the proposed method is equivalent to that of the conventional method. In such a case, the performance of the proposed method can't be verified. In this experiment, only synthesized speech different from that output by the conventional method were used as stimuli for the proposed method.

We prepared as test data 50 synthesized speech pairs. The subjects indicated which of the synthesized speech pairs had the more natural accent. The subjects were five Japanese adults and all were experts on speech perception experimentation. They used headphones to listen to the speech stimuli. The results are presented in Figure 4. The speech output by the proposed methods was selected 68% of the time. This indicates that the proposed methods are effective in creating synthesized speech with natural accents.

5.2. Discussion

The preference test showed that a concatenative speech synthesis system based on the proposed selection method could output synthesized speech that has more correct accenting than a conventional TTS system. We designed two determination functions, which are Function A (simple architecture) and Function B (complex architecture). However, there was no difference in terms of accent validity between the two functions. To construct a better determination function, we may need not only the tonal validity output by the SVM but also additional information.

Since the preference test did not address the total quality of synthesized speech, we conducted another perceptual test, using the Mean Opinion Score (MOS) approach with absolute category rating. As test data, we prepared 40 synthesized speech samples for each method. Other conditions were the same as in the pair comparison test. The results showed that the same MOS score, 3.2, was achieved by the proposed method, using Function A and Function B, and the conventional method. Therefore, while the tonal validity of synthesized speech is improved the overall quality of synthesized speech was unchanged. The reason for this is thought to be degradation due to some unconsidered factor. A preliminary subjective experiment on speech synthesized by the conventional method indicated that a second significant factor in prosody is duration. It is possible to evaluate duration validity of the synthesized speech in the same framework. Accordingly, we plan to research segment selection methods that cover both tonal and duration validity evaluation.

In this paper, we evaluated the proposed method by using Japanese data. However, we think, it's possible that the method will support other languages. To this end, it may be necessary to add SVM features like F_0 transition in the syllable, the type of morpheme and phoneme, and the syllable power, because it's conceivable that which SVM features are effective depends on the language.

6. CONCLUSION

This paper proposed a speech segment selection method that uses a new tonal validity evaluation based on machine learning; its goal to create concatenative speech synthesis systems whose outputs do not exhibit unnatural intonation. Subjective listening tests proved that the proposed method can provide synthesized speech with more natural intonation than the conventional method without degrading the quality of the synthesized speech.

7. REFERENCES

[1] A. Hunt, and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. ICASSP*, pp.369-372, 1996.

[2] T. Hirokawa, "Speech Synthesis Using a Waveform Dictionary," *Proc. Eurospeech*, pp.140-143, 1989.

[3] T. Hirokawa, and K. Hakoda, "Segment Selection and Pitch Modification for High Quality Speech Synthesis using Waveform Segments," *Proc. ICSLP*, pp.337-340, 1990.

[4] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: a new TTS from ATR based on corpus based technologies," *5th ISCA Speech Synthesis Workshop*, pp.179-184, June 2004.

[5] H. Mizuno et al., "Text-to-Speech synthesis Technology Using Corpus-Based Approach," *NTT Technical Review*, vol.2 No.3, March 2004.

[6] M. Chu, and H. Pen, "An objective measure for estimating MOS of synthesized speech," *Proc. Eurospeech*, pp.2087-2090, 2001.

[7] T. Toda, H. Kawai, and M. Tsuzaki, "Optimizing Sub-cost Functions for Segment Selection Based on Perceptual Evaluations in Concatenative Speech Synthesis," *Proc.ICASSP*, pp.I-657-660, May 2004.

[8] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.