Pause insertion prediction using evaluation model of perceptual pause insertion naturalness

Hiroko Muto, Yusuke Ijima, Noboru Miyazaki, Hideyuki Mizuno

NTT Media Intelligence Laboratories, NTT Corporation, Japan

Abstract

This paper describes a pause insertion prediction approach for generating more natural synthesized speech for Text-to-Speech (TTS) synthesis systems. A novel point of the proposed approach is the use of an evaluation model of perceptual pause insertion naturalness in addition to a prediction model based on machine learning. The evaluation model represents the relationship between several features related to pause insertion and the perceptual pause insertion naturalness obtained in a subjective evaluation. First, using a prediction model based on machine learning, we obtain the N-best sequences that indicate whether or not a pause is present at each phrase boundary. We then estimate pause insertion naturalness scores for each N-best sequence using the evaluation model and select the sequence with the highest naturalness score. Objective and subjective evaluation results show that the proposed approach gives better results than a conventional approach.

Index Terms: pause insertion prediction, text-to-speech synthesis, perceptual naturalness, machine learning

1. Introduction

A key to synthesizing speech with improved naturalness is (silent) pause insertion prediction at essential steps in the text-to-speech synthesis process. This is because pause insertions are important for understanding speech. Pause insertions are also important for expressing the characteristics of specific speaking styles. Wang [1] reported that the pause locations of read, expressively read, and spontaneous expressively read speech are different from each other and that irregular pause insertions can improve the expressiveness of synthesized speech. Parlikar [2] reported that natural synthetic speech for the target speaking style can be generated by using a speaking style-dependent pause insertion model. These reports indicate that one needs to be able to predict pause insertions which are suitable for the target speaking style in order to generate expressive synthetic speech.

Machine learning is a suitable way to generate expressive synthetic speech, because it can automatically train the characteristics of pause insertions from training data in the target speaking style. Various machine learning-based approaches have been proposed for predicting pause insertions, such as decision trees [3, 4], Hidden Markov Models (HMMs) [5], Maximum entropy models [6], and Conditional Random Fields (CRFs) [7, 8]. In these studies, various linguistic features, which are extracted from training data, are used for prediction and contribute the prediction performance to improve. On the other hand, it is known that the characteristics of pause insertions have wide variances [9]. Even if the linguistic features are similar, the pause locations have many variations. This indicates it may be difficult to predict more accurate pause insertions by the prediction models trained using only linguistic features.

One useful approach to alleviating this problem would be utilizing not only machine learning-based decisions but also general measures that are independent of specified training data. For example, Kim [10] utilized grammatical rules for statistical pause prediction and showed their effectiveness in reading speaking style. However, this approach would be difficult to apply to a specific speaking style that does not strictly adhere to grammatical rules. To generate expressive synthetic speech, a new measure is needed that can express the characteristics of the target speaking style.

In this paper, we focus on naturalness scores obtained from subjective evaluation. Humans can intuitively judge the naturalness of speech regardless of speaking style. By revealing the features that influence to perceptual pause insertion naturalness and usages of pause insertion prediction, it would be improving the pause insertion performance for specific speaking style. A related study [11] indicates that a perceptual accent naturalness measure obtained from a subjective evaluation is effective at speech segment selection for concatenative speech synthesis. The task of pause insertion prediction would be able to predict more accurate pauses by using perceptual pause insertion naturalness as well as the speech segment selection.

To investigate the characteristics of pause insertions that influence the naturalness of speech, we performed multiple regression analysis to analyze the relationship between several features that related to pause insertions and the naturalness scores obtained by a subjective evaluation using synthesized speech with different pauses. Then, based on the analysis, we propose a pause insertion prediction approach that utilizes the obtained characteristics for machine learning-based pause insertion prediction. We constructed a multiple regression model on the basis of the analysis and used it to evaluate the naturalness of the pause insertions predicted by machine learning. In the proposed approach, we first predict N-best hypotheses of the pause insertion prediction result by using a CRF-based pause insertion prediction. Then, we calculate the naturalness score using the multiple regression model for each N-best hypothesis and rerank them. Finally, the hypothesis with the highest naturalness score is selected from the N-best hypotheses. We conducted a preliminary evaluation of our approach on a reading speaking style corpus and obtained encouraging results.

2. Evaluation of perceptual pause insertion naturalness

In this section, we describe the details of the subjective evaluation using synthesized speech with different pauses and multiple regression analysis to analyze the relationship between several features that related to pause insertions and the naturalness scores obtained by subjective evaluation.

10.21437/SpeechProsody.2014-99

2.1. Subjective evaluation

2.1.1. Speech stimuli

For the subjective evaluation, we used 400 pattern synthesized speech stimuli with different sentences and pause insertion patterns. As a sentence set, 50 sentences were selected from 503 phonetically balanced sentences in the ATR Japanese speech database (Set B). Eight pause insertion patterns were used for each sentence. We did not evaluate all possible pause insertion patterns for each sentence, because it is useless to evaluate patterns humans rarely use. To evaluate patterns that humans use, we utilized patterns from the utterances of several speakers. To evaluate various kinds of pause insertion patterns, we selected speakers whose pause locations are widely distributed. In addition to them, we used the patterns predicted by a rule-based pause insertion predictor [12] to compare with the other patterns. For these purposes, we selected seven speakers and one pause insertion predictor as follows.

- One professional speaker : the pause insertion pattern was used as a general pattern.
- Six non-professional speakers : the average number of their pauses differed significantly. We used them to keep the variation of the pause insertion pattern.
- One Japanese pause insertion predictor : pause insertion pattern was predicted by a manually designed rule.

To evaluate the effects of pause insertion on naturalness, it is desirable to use speech stimuli with the same voice quality and prosody (F0 contour, phoneme duration, accent phrase boundary, and accent type). For this reason, all speech stimuli were generated by a speech synthesizer [13].

2.1.2. Experimental conditions

We had 16 subjects (seven males, nine females) listen to the 400 speech stimuli in random order and rate their naturalness of pause insertion using a 5 point scale (from 1: very unnatural, to 5: very natural). The subjects evaluated each stimulus twice. For each stimulus, the average value of the 16 subjects' scores was used as the naturalness score.

2.1.3. Experimental results

Figure 1 shows the histogram of the naturalness score for each stimulus. 96.5% of the sample scores are over 3 point (fair) because most of speech stimuli employed pause insertion patterns extracted from real utterances, while, the naturalness scores of all sample are not always more than 4 point (natural) and broadly distributed from 2.5 to 4.5. In comparison of the average score for each speaker, the patterns of the professional narrator was the highest and that of the pause insertion predictor was the lowest.

2.2. Multiple Regression analysis

2.2.1. pause distribution features used in analysis

In the multiple regression analysis, we used the following 10 features related to pause insertion, which are called "pause distribution features".

(a) The average and variance values of the number of mora in pause phrases

(b) The average and variance values of the number of content words in pause phrases

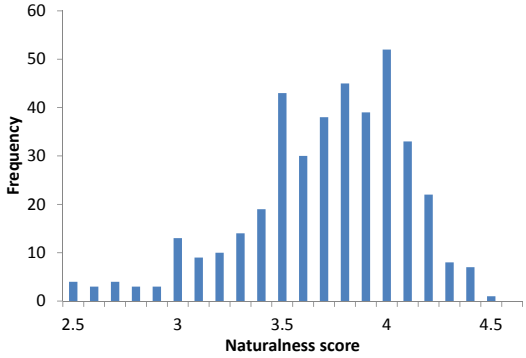

Figure 1: *Naturalness score histogram.*

(c) The existence of four outliers in the number of mora in pause phrases $(\mu - \sigma,\ \mu - 2\sigma,\ \mu + \sigma,\ \mu + 2\sigma)$

(d) The number of pauses inserted between phrases having a dependency relation

(e) The number of non-pauses inserted between phrases not having a dependency relation

These features are calculated for each sentence. A mora is a syllablesized unit in Japanese. "Pause phrase" is one or more consecutive phrases delimited by pauses. (a) and (b) are basic features related to the length and amount of information of each pause phrase. (c) is four features related to outliers. They indicate whether pause phrases that are too long or too short exist. Generally, if such pause phrases exist, the naturalness would become worse. These features can be set to a value of 1 or 0. It is set to 1 if a pause phrase exists in which the number of mora is outside the standard value. In other cases, it is set to 0. The average value μ and the standard deviation value σ, used to determine the standard value, were calculated for all 400 sentences. (d) and (e) are the features related to the relationship of modification between phrases. A previous study [14] showed that these two features affect the naturalness of pause insertions.

2.2.2. Multiple regression analysis results

We performed multiple regression analysis to model the relationship between the pause distribution features given above and the naturalness scores obtained in the subjective evaluation. To investigate the relationship of these features and the naturalness scores, we first calculated a multiple correlation coefficient. We confirmed that the naturalness scores and the estimated ones were correlated; the multiple correlation coefficient was 0.61. We also investigate the effect of each pause distribution feature on the naturalness scores. Partial correlation coefficient for each feature is shown in Table 1. From this table we can see that the features (a) (especially average value), (c) (especially $\mu - 2\sigma$), (d), and (e) were highly correlated for the naturalness score.

3. Proposed approach

By the analysis, we obtain the features that influence to the perceptual pause insertion naturalness. To evaluate the effectiveness of the features for pause insertion prediction, we developed a pause insertion prediction approach that combines the information obtained by the analysis and machine learning-based

Table 1: *Partial correlation coefficients (PCC) for each pause distribution feature.*

pause distribution feature	PCC
Average value of #mora	-0.21
Variance value of #mora	-0.10
Average value of #contents word	-0.03
Variance value of #contents word	-0.01
Outlier ($\mu - \sigma$)	-0.08
Outlier ($\mu - 2\sigma$)	-0.23
Outlier ($\mu + \sigma$)	-0.02
Outlier ($\mu + 2\sigma$)	-0.06
#pauses inserted between dependency relation phrases	-0.42
#non-pauses inserted between non dependency relation phrases	-0.24

pause insertion prediction. We constructed a multiple regression model on the basis of the analysis and used it to evaluate the naturalness of the pause insertions predicted by machine learning. We call it the "evaluation model". A block diagram of the proposed approach is shown in Fig. 2. It consists of two stages: a prediction stage and an evaluation stage.

In the prediction stage, we predict a Boolean sequence that indicates whether a pause has to be inserted after each phrase within the input phrase sequence. (We call this Boolean sequence a "pause assignment sequence"). A CRF-based pause insertion prediction model is used for prediction and N-best pause assignment sequences are output.

In the evaluation stage, for each N-best sequence, we extract pause distribution features. We then estimate the naturalness score using the extracted pause distribution features and an evaluation model of perceptual pause insertion naturalness. Finally, we select the pause assignment sequence with the highest naturalness score among the N-best sequences.

Although pause insertions were mainly predicted by word units in previous studies, however, in this study, we predict them by phrase units (also called "bunsetu" [15]) because pauses are usually inserted into phrase boundaries.

4. Pause insertion prediction model

4.1. Pause insertion modeling with CRF

In this subsection we describe the pause insertion prediction model. Whether a pause should be inserted after a phrase greatly depends not only on the various linguistic features of the phrase but also on what comes before and after the pause. Machine learning approaches may be effective for modeling such comprehensive and complex features. In this study we used the CRF, which is an effective machine learning approach for sequential labeling. Given a sequence of vectors of the linguistic features $\boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M)$ with M phrases, the goal is to find a highest probability sequence of pause assignment labels $\boldsymbol{y} = (y_1, \cdots, y_M)$ for M junctures after every phrase. The linguistic features of sequence $\boldsymbol{x}$ are described in Sect. 4.2. Each y_i, the i-th pause assignment label in sequence $\boldsymbol{y}$, can be 1 or 0. If a pause is inserted immediately after the i-th phrase, y_i is 1. In other case, y_i is 0.

In this study we used the linear chain CRF, a special form of CRF. The conditional probability $P(\boldsymbol{y}|\boldsymbol{x})$ is calculated as

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp(\boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}} \exp(\boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{y}))}$$

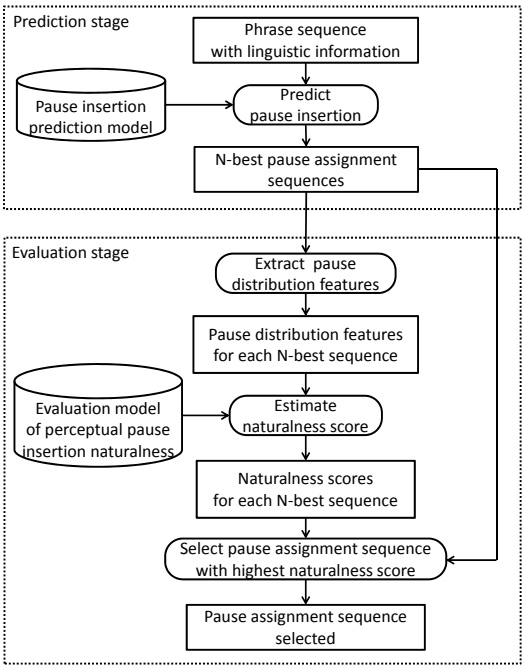

Figure 2: *Overview of proposed approach.*

where $\boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{y})$ is a feature function and $\boldsymbol{w}$ is a parameter to be estimated from training data. CRF are usually trained by maximizing the log-likelihood over a given training set. In this study we obtained the N-best sequences in order of the probability of $P(\boldsymbol{y}|\boldsymbol{x})$.

4.2. Linguistic features for pause insertion modeling

These linguistic features were used for pause insertion modeling .

(1) Text of current phrase
(2) Syntactic category of current phrase
(3) Flag of whether current phrase modifies next phrase
(4) Text of last word of current phrase
(5) Part of speech (POS) of last word of current phrase
(6) Pronunciation of last word of current phrase
(7) Text of first word of next phrase
(8) POS of first word of next phrase
(9) Pronunciation of first word of next phrase
(10) Above features' quin-context and combinations of them

In this study, we modified the word-based features used in conventional approaches [7, 8] to suitable ones for the prediction by phrase units. We also used the features that indicate the dependency relationships between phrases.

5. Evaluation

5.1. Experimental conditions

In the following experiment, a Japanese news corpus was used. It contained 1000 Japanese news sentences uttered by one professional male narrator, with an average of 70.0 words per sentence, 23.4 phrases and 11.7 pauses. The POS, pronunciation,

Table 2: *The results of pause insertion prediction.*

	Recall	Precision	F-measure
Proposed	93.3	72.4	81.5
Conventional	93.6	68.4	79.0

and dependency relation were labeled automatically using a Japanese morphological analyzer [16] and a Japanese dependency parser [17]. Pause locations were labeled manually according to the speech data.

As a conventional approach, the 1-best result predicted by CRF-based pause insertion prediction model was used. CRF++ toolkit [18] was used for pause insertion prediction model training. We set the parameter N (the number of N-best sequences) as 10 from the preliminary experiment results.

5.2. Objective evaluation

We first compared the prediction performance of the proposed approach with that of the conventional approach. From the corpus, 100 sentences were used as the evaluation data and the rest were used as training data for the pause insertion prediction model. We performed a 10-fold cross validation test. The prediction performance was evaluated by precision, recall, and F-measure. Where, precision means the ratio of correctly inserted pauses to the total number of inserted pauses, and recall means the ratio of correctly inserted pauses to the total number of pauses in the evaluation data.

The results are shown in Table 2; the proposed approach had slightly lower recall but higher precision and F-measure than the conventional approach.

5.3. Subjective evaluation

We also compared the naturalness of synthesized speech generated from the results predicted by each approach by a pair comparison test. Since the sentences in the corpus were too long to evaluate subjectively, we truncated them to short sentences that can be listened to easily. From the corpus we selected 30 sentences for evaluation and truncated them to 148 short sentences, averaging eight seconds in length. The rest were used for training the pause insertion prediction model. We randomly selected 30 short sentences from the short sentences in which pause location differences were found between results predicted by the proposed and the conventional approaches. The synthesized speech was generated by the same speech synthesizer used for the subjective evaluation described in Sect. 2.1. The pause durations in the synthesized speech were consistently 0.5 seconds. Subjects were 24 persons (12 males, 12 females), and presented a pair of synthesized speech samples in random order and then asked which samples had better naturalness.

The preference scores obtained for the experiment (Fig. 3) show the proposed approach outperformed the conventional approach. This shows that using an evaluation model of perceptual pause insertion naturalness is an effective way to improve pause insertion naturalness in synthesized speech.

We conducted the objective evaluation using the short sentences used for the subjective evaluation. The prediction performance was shown in Table 3. From this table we can see that the prediction performance was almost the same as those using the original sentences.

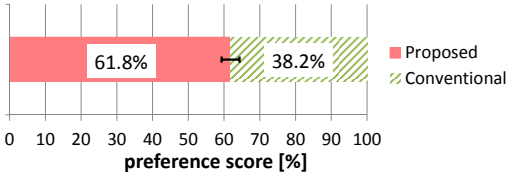

Figure 3: *Preference score from the subjective evaluation. (Error bars show the 95% confidence intervals.)*

Table 3: *The results of pause insertion prediction. (Using short sentences used for subjective evaluation)*

	Recall	Precision	F-measure
Proposed	94.1	74.4	83.1
Conventional	95.2	72.5	82.3

6. Conclusion

In this paper, we proposed a pause insertion prediction approach involving the use of an evaluation model of perceptual pause insertion naturalness for generating natural synthesized speech. Objective and subjective evaluations demonstrated the proposed approach provides better results than the conventional approach. This shows that consideration of perceptual pause insertion naturalness to pause insertion prediction is an effective way to improve naturalness in synthesized speech.

Currently the proposed approach's effectiveness has been demonstrated only for the reading speaking style. A future task is to explore its effectiveness for other speaking styles. We also intend to analyze pause duration naturalness and attempt to develop a pause duration estimation approach for generating more natural synthesized speech.

7. References

[1] X. Wang, A. LI, C. Yuan, " A Preliminary Study on Silent Pauses in Mandarin Expressive Speech," Speech Prosody, pp. 673–676, 2008.

[2] Parlikar. A and A.W. Black, " A Grammar Based Approach to Style Specific Phrase Prediction," INTERSPEECH, pp. 2149–2152, 2011.

[3] M. Q. Wang and J. Hirschberg, " Automatic classification of intonational phrase boundaries," Computer Speech and Language, vol. 6, pp. 175–196, 1992.

[4] P. Koehn, S. Abney, J. Hirschberg and M. Collins, " Improving intonational phrasing with syntactic information," ICASSP, pp. 1289–1290, 2000.

[5] P. Taylor and A. W. Black, " Assigning phrase breaks from part of speech sequences," Computer Speech and Language, Vol. 12, pp. 99–117, 1998.

[6] J. Li, G. Hu and R. Wang, " Chinese prosody phrase break prediction based on maximum entropy model," INTERSPEECH, pp. 729–732, 2004.

[7] Y. Qian, Z. Wu, X. Ma and F. Soong, " Automatic prosody prediction and detection with Conditional Random Field (CRF) models," ISCSLP, pp. 135–138, 2010.

[8] J. Sun, J. Yang, J. Zhang and Y. Yan, " Chinese prosody structure prediction based on Conditional Random Fields," 5th International Conference on Natural Computation, pp. 602–606, 2009.

[9] H. Fujisaki, S. Ohno and S. Yamada, " Analysis of occurrence of pauses and their durations in Japanese text reading," ICSLP, Vol. 4, pp. 1387–1390, 1998.

[10] B. Kim and G. Lee. "Statistical/Rulebased Hybrid Phrase Break Detection," ICSP, 1999.

[11] A. Yoshida, H. Mizuno and K. Mano, " Segment selection method based on tonal validity evaluation using machine learning for concatenative speech synthesis," ICASSP, pp. 4617–4620, 2008.

[12] K. Matsuoka, E. Takeishi and H. Asano, " AUDIOTEX. A Text-To-Speech System for Japanese Text", 52nd National Convention of IPSJ, Vol. 2, pp. 409–410, 1996. (in Japanese)

[13] K. Mano, H. Mizuno, H. Nakajima, N. Miyazaki and A. Yoshida, " Cralinet—Text-To-Speech System Providing Natural Voice Responses to Customers," NTT Technical Review, Vol. 18, No. 11, pp. 19–22, 2006.

[14] N. Kaiki and Y. Sagisaka, " Study of pause insertion rules based on local phrase dependency structure," IEICE, Vol. J79-D-II, No. 9, pp. 1455–1463, 1996.

[15] Y. Zhang and K. Ozeki, " Automatic bunsetsu segmentation of Japanese sentences using a classification tree," Language Information and Computation, pp. 230-235, 1998.

[16] K. Imamura, K. Saito and H. Asano, " Basic Japanese text analysis technology as a platform for Knowledge extraction," NTT Technical Review, Vol. 6, No. 9, 2008.

[17] K. Imamura, G. Kikui and N. Yasuda, " Japanese Dependency Parsing Using Sequential Labeling for Semi-spoken Language," ACL, pp. 225–228, 2007.

[18] CRF++: Yet Another CRF toolkit , http://crfpp.sourceforge.net/.