Analysis of major factors of naturalness degradation in concatenative synthesis

Toshio Hirai† Hisashi Kawai†,‡ Minoru Tsuzaki†,* Nobuyuki Nishizawa†

†ATR Spoken Language Translation Research Labs., 2–2–2 Hikaridai Seika-cho Souraku-gun Kyoto 619–0288, Japan
‡KDDI R&D Labs., Japan *Kyoto City University of Arts, Japan
toshio.hirai@atr.jp

Abstract
To effectively improve a speech synthesis system, it is important to find and focus on improving the modules whose effect on naturalness degradation in synthesized speech are the largest. In this paper, we describe the design of a perception experiment to measure the effect of each module separately. Synthesized speech stimuli whose intermediate information is modified during a synthesis process are used in the experiment. A perception experiment in which a Japanese concatenative speech synthesis system was evaluated revealed that the text processing module and a part of the feature prediction module (for the fundamental frequency) of the system were the major factors in degrading naturalness.

1. Introduction
Concatenative Text-to-Speech (TTS) systems are composed of 4 major elemental modules, namely a text processing module, an acoustic- and prosodic-feature prediction module, a segment selection module, and a waveform concatenation module. Although all the modules are expected to be improved, the resources to achieve this are not always sufficient. In such a case, it is necessary to find and focus on improving the modules whose effect on naturalness degradation in synthesized speech are the largest. In order to find such modules, the degradation amount must be measured separately. However, only the gross system performance has been focused on (e.g., intelligibility/understandability tests for syllable/word/sentence, and a prosodic naturalness evaluation test) in past research related to Japanese speech synthesis systems[1, 2, 3].

In this paper, we describe the design of an experiment for the evaluation of each module’s performance separately. The target modules for evaluation are limited to the text processing module and the feature prediction module (to put it more precisely, its subordinate parts, i.e., acoustic- and prosodic-feature prediction part), whose performance is closely related to the naturalness of synthesized speech. In the case of the performance evaluation of the text processing module, the amount of naturalness degradation of a speech stimulus synthesized with non-corrected intermediate output from the module is measured with reference to that of the stimulus synthesized with phone- and prosody-corrected intermediate output. The amount of naturalness degradation of the stimulus synthesized with the features which are generated from one of the estimated features and the features extracted from natural speech is measured with reference to the stimulus synthesized with the features extracted from natural speech in the case of a performance evaluation for the feature prediction module. We applied the method to a concatenative speech synthesis system called XIMERA[4], which has been developed at ATR. The target languages of XIMERA are Japanese and Chinese, although the method was applied only for Japanese in this paper. An experiment in which 40 listeners participated was conducted, and the details of the results and a discussion of the experiments are described.

2. XIMERA speech synthesis system
The data flow in XIMERA is as described below: (1) The text processing module processes the input Japanese text (a mixture of Chinese characters and Japanese phonetic characters) in order to note linguistic- (morphologic and phonetic) and prosodic- (accentual) information. (2) The passed information is used to predict the time series of acoustic/prosodic information with an acoustic- and prosodic-feature prediction module. In the case of XIMERA, HTS (HMM-based Triple S (Speech Synthesis System))[5, 6] is used as the module. For the training of the feature prediction model in HTS, sentences in an ATR phonetically balanced task (ATR503)[7] (503 sentences), sentences in a travel guidebook task (665), and sentences in a newspaper story task (498) were used (total: 1,666). The set of the predicted features is called the “target.” The acoustic feature is Mel-cepstrum (mcep), and the prosodic features are segment duration and speech fundamental frequency (∫0). (3) In the segment selection module, “target cost” and “join cost” are calculated based on the target features and the attributes of speech segments in a speech database. The costs are integrated into one cost for each segment sequence candidate, and an appropriate sequence that shows the lowest cost is selected. (4) Finally, the selected segments are joined smoothly into a waveform in the waveform generation module, and it is output as synthesized speech. In the following section, a method to estimate the naturalness degradation of synthesized speech caused by modules, and its application to the modules in XIMERA are detailed.

3. Analysis of major factors of naturalness degradation
The target that is used to generate a “natural enough” synthesized speech (hereafter called “speech (a)” or just “(a)”) is the “natural target.” There are many candidates for the natural target. A target that is extracted from natural speech “(n)” is a good choice since it is easy to acquire and its stability is guaranteed. In this paper, the target is treated as the natural one. The influence on naturalness degradation of synthesized speech caused by the feature prediction module’s imperfection is estimated separately by measuring the naturalness degradation of speech that is synthesized with the merged features of the natural target and a predicted feature (either mcep, segment duration, or ∫0) from speech (a). (Each synthesized speech is
Figure 1: Illustration of the speech stimulus categories and target used for synthesis.

(1) the text processing module, (2) the feature prediction module, and (3, 4): the segment selection module and the waveform generation module (Refer to Section 2.)

*: features predicted from the corrected intermediate output from the text processing module.

Table 1: Correctness in text processing

<table>
<thead>
<tr>
<th></th>
<th>no corr.</th>
<th>corr. req.</th>
</tr>
</thead>
<tbody>
<tr>
<td>sent. num.</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>mora num.</td>
<td>1,007</td>
<td>1,248</td>
</tr>
<tr>
<td>phon. desc.*</td>
<td>(100, 100)</td>
<td>95.0, 94.9</td>
</tr>
<tr>
<td>acc. boundary*</td>
<td>93.8, 88.1</td>
<td>76.9, 74.2</td>
</tr>
<tr>
<td>acc. phrase num.†</td>
<td>215</td>
<td>213</td>
</tr>
<tr>
<td>acc. type‡</td>
<td>87.0</td>
<td>91.5</td>
</tr>
</tbody>
</table>

*: correspondent rate (= H/total × 100) and accuracy (= (H – 1)/total × 100) in percentage (H: hit, F: insertion), †: the number of active accentual phrases (phonetic description and boundary correspond to the natural speech), ‡: accent type correspondent rate (see the definition of ‘*’). In the calculation of the correspondent rate and accuracy for boundary, the post boundary was not treated as an H one if the phonetic description differed from the natural speech.

4. Experiment

4.1. Sentence selection

As described in Section 3, the speech (synthesized or natural) that is used in the subjective experiment for naturalness evaluation must satisfy a condition: the phoneme sequence of (b–corr) must be the same as that of the corresponding natural speech following correction of the intermediate output from the text processing module. In this study, the allowed correction items were: accentual phrase boundary, accent type of accentual phrase, and phonetic description. The forbidden correction items were: insertion/deletion of pauses and voiced/unvoiced sound changing. The sentences were selected from the following tasks: novels, ATR503, travel guidebooks, and newspaper stories.

Originally, it was attempted to make the number of sentences equal in each task. However, since the number of sentences that satisfy the conditions becomes small in the case of long sentences (like those included in novels and newspaper stories), the number of sentences varied widely depending on the tasks. Therefore, we decided not to control the number. Finally, the number of sentences were 9, 30, 39, and 20 for the novel, ATR503, travel guidebooks, and newspaper story tasks, respectively (total: 98). In each task, the number of sentences that did not require correction of the phonetic description were balanced with those that required correction. The correctness of the text processing in the selected sentences is shown in Table 1. The mean and standard deviation (S.D.) of the sentence length were 23.0 and 12.1 (mora/sentence). The following are some correction examples. Phonetic description: input "cyano no (a big roof)," pre-correction "dayaneno," accentual phrase boundary: input "boNyarihiteteita (was dazed)," pre-correction "boNyariahitetita," accent type: input "hizaka’keda (be a lap robe, type 3)," pre-correction "hizakake’d (type 4)."

4.2. Synthesis conditions

In speech synthesis, a 47-hour female speech database recorded at ATR was used. The speech whose transcription was the same as the synthesized speech was eliminated from the database during the synthesis processing. Speech data in the database and the synthesized speech were 16 kHz sampled and quantized with 16 bits. The frame shifts were: 5 ms for mcep and predicted F₀, and 10 ms for F₀ in the natural target. This made the F₀ sampling frequency different in different speech categories; 10 ms: (a), (a–mcep), and (a–dur), 5 ms: (a–F₀), (b), and (b–corr).

The 98 selected sentences were synthesized in the 6 methods described in Figure 1. Natural speech of randomly selected 12 sentences were additionally presented to each listener. As a result, 600 speech stimuli (= 98 × 6 + 12) for each listener were prepared. The stimuli were shuffled randomly on the condition that the same sentence does not appear within the adjacent ± 12 sentences.
The phonetic boundary information and $F_0$ data of ATR503 were corrected manually. However, the other tasks’ information/data were in their natural state (as detected automatically). For $F_0$ extraction, STRAIGHT TEMPO was used[8].

4.3. Listeners

The listeners consisted of 40 persons (male: 19, female: 21). They lived in the Tokyo area until they were 12 years old. They had no hearing problems. Each group of 20-, 30-, and 40-year-old listeners was balanced in gender.

4.4. Sound equipment

The experiment was conducted in a quiet room. Notebook type personal computers were used in order to play back the synthesized/natural speech with software for hearing experiments. MDR-Z900 (SONY Inc.) headphones were used in the experiment.

In order to achieve a uniform sound pressure level in each device, the playback systems were calibrated every day with the following method. The voltage (AC) of the left channel of the headphones was set to 0.133 V with a 1-kHz sine waveform (16 kHz sampled, maximum amplitude was ~4.29 dBFS (Full Scale) = $20 \log_{10}(20000/2^{15})$). The mean sound pressure level of 20 randomly selected speech stimuli was 81.5 (dB(A)) with S.D. 1.36.

4.5. Evaluation procedure

The listeners were asked to rank the naturalness of each stimulus in 5 levels: “1” means “unnatural” and “5” means “equal to natural speech.” They were allowed to listen to the same speech as many times as they wanted. It was easy to distinguish a natural speech in stimuli, so the listeners were asked to judge regardless of whether they noticed it was the natural one or not.

For the practice, 10 speech stimuli were selected randomly from the synthesized speech, and they were listened to before the stimuli. The stimuli were separated into 4 parts (150 speech stimuli each), and the listeners were forced to take a 2–5 minute rest between the intervals. They were also allowed to take a short break at any time they wanted. The mean time required to complete the experiment by a listener was about 2 hours including breaks.

5. Results

In advance, listeners whose evaluation was biased (and who did not use the full range of evaluation) were sought out and omitted in order to remove their influence. First, the listeners whose judgments for natural speech were lower than ‘5 (equal to natural speech)’ were sought out since their evaluation might be too “strict.” As a result, 3 listeners fit the condition (male: 2, female: 1). Next, listeners who had never judged ‘1 (unnatural)’ for synthesized speech were sought out since their evaluation might be too “loose.” In this screening, 4 listeners were found (male: 1, female: 3). In the following analysis, the judgments by these 7 listeners are omitted.

The bar graph of the mean of evaluation for each speech category is shown in Figure 2. The S.D. of the evaluation result in each speech category varied from 0.71 (n) to 1.22 (b). The reason why the deviation of (b) was largest is that the stimuli included both sentences that require correction on the phonetic description (mean: 2.79, at the bottom of the white arrow next to the (b–corr) bar) and others (3.07). The naturalness difference between these 2 groups in (b–corr) was small (3.26 vs. 3.31). Tukey-Kramer’s honest significant difference method[9] revealed statistically significant differences between the means of (a-F0) and each of the other conditions, i.e., (a), (a-mcep), and (a-dur) ($p < 0.01$). Additionally, it was also revealed that the means of (b-corr) and (b) were significantly different (Welch’s two sample t-test, $p < 0.01$).

The relationship between the amount of phone-/prosody-correction and the evaluation improvement from (b) to (b–corr) was surveyed. The considered correction items are shown at the left side of Table 2. Only the active accentual phrase that is defined in Table 1 was considered for the accent type error. The relationship between these items and the evaluation improvement in each speech stimulus was analyzed with multiple linear regression. The regression coefficients are shown at the right side of Table 2. The multiple correlation coefficient was 0.584. An investigation showed that there can be two reasons why there were two negative coefficients: (1) the correction number was small (3 and 13 respectively), or (2) there were some items that coincidentally co-occurred with them and the items behaved as canceling factors.

For reference on the performance of the feature prediction module, the root mean square (RMS) of the prediction error for duration and $F_0$ are shown in Table 3. The values in the “Closed” and “Open” lines are calculated from the sentences...
Table 3: RMS of prediction error for duration and F0

<table>
<thead>
<tr>
<th></th>
<th>Duration (ms)</th>
<th>F0 (semitone)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closed</td>
<td>19.2</td>
<td>1.81</td>
</tr>
<tr>
<td>Open</td>
<td>20.4</td>
<td>2.04</td>
</tr>
</tbody>
</table>

which were used as the training data for HTS or from the other sentences, and the values in the "Duration" and "F0" columns are calculated from the target for (a–dur) (between all phoneme pairs) or from that for (a–F0) (between all effective F0 pairs).

The RMS difference between the natural target and the actual duration or F0 in (a) (5 prediction points in each phoneme) was 10.4 (ms) or 1.23 (semitones). Since the meaning of the prediction error of mcep is not clear, the error was not surveyed.

6. Discussion

From Figure 2, it is clear that:

- the order of the influence on degradation of naturalness between the merging manipulations was: F0 > duration > mcep.
- the interval between (n) and (a) is roughly equal to that between (a) and (b), and
- the interval between (a) and (b–corr) is roughly equal to that between (b–corr) and (b).

Since the correction for the text processing result and the merging of the predicted F0 with the natural target were major factors for the naturalness degradation of synthesized speech, the improvement of the text processing module and the F0 prediction part in the feature prediction module would be directly linked to an improvement of the synthesis system.

The naturalness difference between (n) and (a) would be caused by the limitations of the segment selection module, the limitations of the acoustic-/prosodic-feature’s ability to represent speech, or the insufficient size of the speech database for segment selection. The difference would be expected from the fact that the values of the RMS difference between the natural target and the actual features in (a) were not negligible compared with the values in Table 3.

There were 28 stimuli whose naturalness for (b–corr) degraded from (b). In this paragraph, one of 2 stimuli which degraded more than one level is examined. In the example, there was only one text processing error in (b): one diphthong was corrected into a long vowel ("seifu→"seifu (government)"). As a result of the correction, the mean naturalness degraded from 4.00 to 2.94. The direct cause for the degradation would be the unnatural F0 pattern at the accentual phrase located 3 phrases after the corrected vowel. There was no difference between the predicted F0 in (b) and (b–corr) around the accentual phrase whose F0 pattern change in synthesized speech caused the degradation of (b–corr). Such a phenomenon occurs because the segment selection module occasionally selects completely different segments after some phrases from a different target are presented though the difference of the target is small.

The effect of the auto-segmentation or auto F0 extraction in the tasks except for ATR503 seems to be small since not all of the naturalness of ATR503 was highest. The naturalness of ATR503 for (a) and (a–*) shows only a small (or no) improvement from (b–corr). This may occur because the task was used for the HMM training for feature prediction in HTS and the prediction result of (b–corr) was nearly equal to the training data of (a).

As mentioned in Section 3, the features extracted from a natural speech waveform used in this study were just one candidate for the "natural target." The evaluation result would be different if another kind of "natural target" was used. In order to evaluate the naturalness degradation degree with precision, various kinds of "natural targets" should be used for the merging manipulation, and the average of the evaluation should be taken in each category.

7. Summary

A method to estimate the naturalness degradation caused by a lack of ability in the elemental modules that compose a speech synthesis system is described in this paper. In the method, the naturalness degradation or improvement of speech that is synthesized by a controlled target from speech that is synthesized by a natural target (which shows the best naturalness) or from synthesized speech that is in an as-is status (which shows the worst naturalness) is evaluated subjectively. As the controlled target, one of the predicted features is merged with the natural target, or the target predicted from the corrected intermediate output from the text processing module is used. We applied the method to the XIMERA speech synthesis system, which was developed at ATR. The result of a perception experiment in which 40 listeners participated shows that the improvement of the text processing module and the F0 prediction part in the feature prediction module should be given a higher priority in system development.

8. Acknowledgments

This research was supported in part by the National Institute of Information and Communications Technology. The authors are thankful to Dr. Hiroaki Kato, ATR Human Information Science Laboratories, for his comments on the listener selection.

9. References


