SSW6

Utilization of an HMM-Based Feature Generation Module in 5 ms Segment Concatenative Speech Synthesis

Toshio Hirai† *Junichi Yamagishi*‡ *Seiichi Tenpaku*†

† Arcadia, Inc., Japan
‡ The Centre for Speech Technology Research, University of Edinburgh, UK
thirai@arcadia.co.jp

Abstract

If a concatenative speech synthesis system uses more short speech segments, it increases the potential to generate natural speech because the concatenation variation becomes greater. Recently, a synthesis approach was proposed in which very short (5 ms) segments are used. In this paper, an implementation of an HMM-based feature generation module into a very short segment concatenative synthesis system that has the advantage of modularity and a synthesis experiment are described.

1. Introduction

Speech synthesis is a technique for converting text into speech. Currently, concatenative speech synthesis systems are entering the mainstream because they can achieve high quality speech without great difficulty. In a concatenative speech synthesis system, an amount of recorded speech samples and their features are stored in a database ("corpus.") At synthesis, an appropriate speech segment sequence is selected from the corpus and concatenated smoothly. The selection is executed according to the feature time series (target) that is generated from input text. A concatenative speech synthesis system increases its potential to generate natural speech if the system uses shorter speech segments, because the concatenation variation becomes greater. Recently, a synthesis approach was proposed in which very short segments (5 ms) were used [1, 2, 3, 4].

We have been trying to improve the naturalness of the synthesized speech. At the same time, we sought the feature generation module since our system lacked the module [1]. Another 5 ms segment synthesis method proposed by Ling et al. uses HTS (HMM-based Triple S (speech synthesis system)) [5] as a feature generation module [2]. HTS has the ability to generate a time series of vectors (mel-cepstrum or "mcep") from Linguistic-/Prosodic-Information(LPI) that is produced from input text. In our system, HTS is also adopted as the feature generation module.

The method proposed by Ling et al. requires the time series of the mean and variance of features as the target value for synthesis. Since it is difficult to get paired information which guarantees "natural" synthesized speech, it is also difficult to isolate the cause (a problem of the feature generation module *or* of the feature-to-speech module) when the synthesized speech is not natural enough. On the other hand, in our method, only the mean value is required as the target to synthesize. It would be an advantage since if the feature (which corresponds to the mean value and must have the information to synthesize natural speech) extracted from natural recorded speech is used as the target in our method but the synthesized speech is not natural enough, it can be said that the cause rests with the feature-to-speech module.

This paper is organized as follows. The next section (Section 2) introduces the processing outline in the 5 ms segment concatenative speech synthesis including the utilization of the feature generation module in HTS. The following two sections (Sections 3 and 4) present a synthesis experiment and its results, in which 450 Japanese utterances were used. Section 5 discusses the findings in the experiment, and Section 6 summarizes this paper.

2. Concatenative speech synthesis using 5 ms segments with an HMM-based feature generation module

2.1. Analysis stage

2.1.1. Corpus construction

Speech data are analyzed every 5 ms, and extracted features are stored in a speech corpus during the analysis stage for corpus construction. The extracted features are the speech fundamental frequency (F_0), power, and spectrum. As the spectral information, mcep is adopted. These features are used to get the target cost in the synthesis stage. To construct a corpus automatically and to avoid contamination of the F_0 extraction errors, it is not the scalar F_0 value (as in conventional speech synthesis systems including Ling's method [2]) but the lower frequency part of a power spectrum (the upper frequency bound is set by the highest F_0 of the speech data) which is treated as the F_0 information in our method. Hereafter, the lower frequency part of a power spectrum is also denoted as "F_0 information." **Figure 1** shows an example of the F_0 information. As can be seen from the figure, the "ridge" of the power spectrum contour corresponds to the scalar F_0 pattern ('+') very well.

In the synthesis stage, FFT-based power spectrum is used for the calculation of concatenation distortion at the concatenation point. Therefore, FFT analysis centered at the ends of 5 ms segments is also executed. The processing flow is illustrated in **Figure 2**.

2.1.2. Feature generation module building

In HTS, the relationship between LPI and features (F_0 and mcep) is analyzed in order to construct a feature generation model. In our method, it is required to generate F_0 information by the model. In this report, the F_0 information is merged into the mcep and analyzed as one stream in model training for simplicity.

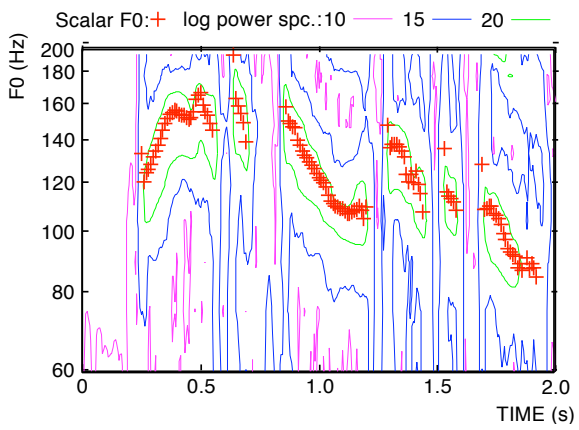


Figure 1: *Contour of the lower frequency part of power spectrum with scalar F_0.*

Initial part of sentence J21 in ATR503: "nyu:gakushIkeNo ukeru tokiyori (than an entrance examination)"

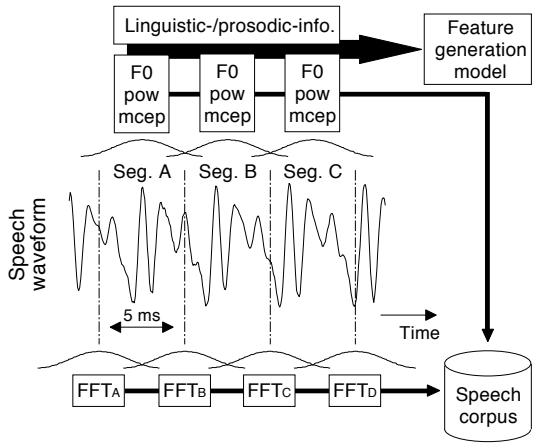
2.2. Synthesis stage

2.2.1. Feature generation

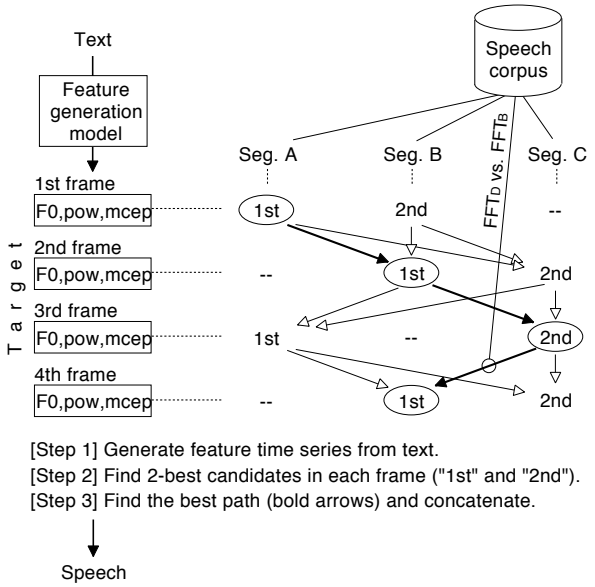
In the synthesis stage, input text is used to generate the feature time series by HMMs trained in the analysis stage as target vectors that are used to synthesize the required speech sound. The generated feature is decomposed into F_0 information, power, and mcep.

2.2.2. Feature-to-speech processing

Speech segments similar to the generated feature are searched for in each segment in the speech corpus using a target cost function, and the N-best segments are selected as candidates in each frame. Next, all combinations of the candidate connections are evaluated, and the segment sequence exhibiting the lowest connection distortion per concatenation point is concatenated in order to generate synthesized speech.



(a) Analysis stage



(b) Synthesis stage

Figure 2: *Processing flowchart of analysis/synthesis stages.*

3. Experiment

3.1. Speech material

We used a phonetically balanced Japanese speech database ("ATR 503 sentences [6]," ATR503) spoken by a male speaker that is attached to HTS [5]. Speech data in the database were sampled at 16 kHz and quantized with 16 bits. For corpus construction and feature generation module building by HTS in the analysis stage, we used the 450 utterances of the database (the groups A, B, ..., and I). The total number of 5 ms segments was 481,207.

3.2. Analysis conditions

These are the segment analysis conditions: frame length, 1,024 points (64 ms) for F_0 analysis and 512 points (32 ms) for power and mcep analysis; Hanning windowing; frame step width, 80 points (5 ms); F_0 analysis, from 1st to 19th orders of the power spectrum (the frequency of the 19th channel is 296.875 Hz ($= 16,000/1,024 \times 19$)); the order and α in mcep analysis, 24 and 0.42. The zero-th term of mcep was not used as the power. Instead, power was calculated from the windowed signal directly. For the extraction of the mcep parameter, we used the "mcep" command in Speech Signal Processing Toolkit [7].

In paper [1], the lower frequency part of a power spectrum was normalized in each segment in order to eliminate power information by dividing each original value by the summation. However, if the normalized value is used in the HTS training, the generated feature sometimes shows negative values though it should be positive. To avoid this problem, the normalization was not executed in this report. Because the velocity ("Δ") and the acceleration ("Δ^2") of the value are considered in parameter generation of HTS [8], the smoothness of them was ensured.

These are the analysis conditions for distortion measurement at the segment edge: Frame length was 256 points (= 16 ms) with 0.97 pre-emphasis and Hanning windowing.

3.3. HTS processing conditions

The version of HTS was 2.0 [5]. From the remaining 53 sentences that were not used for model building, five were chosen for a synthesis experiment. Scalar F_0 information was also included in the training features since the training became unstable without it. The feature sequences were directly

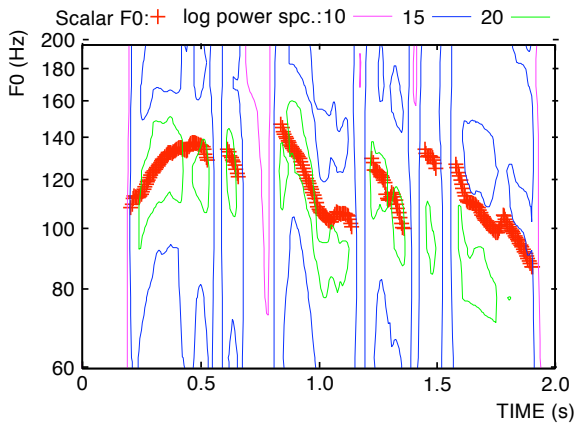


Figure 3: *Parameters generated by HTS.* Refer to the note of Figure 1.

generated from HMMs that was trained with HTS. The HMMs are 5-state left-to-right context-dependent HMMs, and each state has 2 Gaussian probability density functions. We utilized an EM-based iterative parameter generation algorithm which is detailed in Section 2.3 of [8] (Case 3) and the information is stored in the directory "gen/qst001/ver1/2mix/2" in the HTS system.

3.4. Synthesis conditions

N in N-best was set at 300. This is the procedure to calculate the distance between a target segment and a segment in the corpus: (1) For all features (F_0, power, and mcep), execute (1–1) and (1–2). (1–1) For each target segment, calculate the Euclid distance of the feature from all segments in the corpus. (1–2) Each distance is normalized by the mean and standard deviation of all distances. (2) The summation of weighted previous-/current-/post-positional distances of all features is treated as the definitive distance. The weights for previous-/current-/post-position were 1, 3, and 1.

For the distortion measure at concatenation points, the Kullback-Leibler distance of the FFT-based power spectra [9] with the consideration of the powers was adopted. The Dijkstra's shortest path search algorithm [10] was used for the full path search. Consideration for previous-/post-target distance and concatenation distortion ensures the smoothness of synthesized speech indirectly. Finally, the cross-fade technique for frame sized segment concatenation was used to generate the speech waveform.

4. Results

It took about 7 hours to complete the learning of HMMs by a computer with a 2.4 GHz CPU and 768 MB memory, excluding the feature extraction time from recorded speech. A sample of the generated F_0 information is shown in **Figure 3** with the generated scalar F_0. For 5 ms segment synthesis, it took about 1.6 hours for a sentence. The spectrum, F_0, etc. of the synthesized speech are shown in **Figure 4** with those of the recorded speech. Synthesized speech samples can be listened to at: http://www.arcadia.co.jp/~thirai/ssw6.

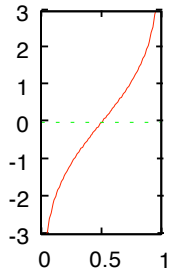


Figure 5: *An example of a finite-infinite transformation (equation (1)).*

5. Discussion

The speech synthesized by the proposed method seems to have better voice quality (speaker's individual reproducibility) than that generated by the original HTS. In 5 ms segment speech synthesis, sometimes a buzz noise appears (e.g. /N/ in "he:kiN" of speech sample J11). It might be caused by the monotonous target value in a part where a segment sequence is used repeatedly in the part in synthesized speech. It would be suppressed by adding a small random noise to the target feature.

The intonation of the speech synthesized by the 5 ms segment synthesis and by the HTS has problems. For example, the intonation pattern of 5 morae accentual phrase "unagiyani (at an eel restaurant)" of sample J01 should be "LHHHH," but the patterns realized in the synthesized speech were "LHHHL." (H: High, L: Low.) It might be caused by the mismatch between recorded speech and the LPI of it in the training data. Therefore, if LPI is corrected according to the recorded speech, the intonation quality of synthesized speech would be improved.

As mentioned in section 3.2, the normalization of the lower frequency part of a power spectrum was not executed in this report. In order to suppress the appearance of the negative value in a generated feature, it would be effective to use a function that transforms finite-domain $[0, 1]$ to infinite-range $(-\infty, \infty)$, such as the logistic transformation:

$$y = \log \frac{x}{1-x}, \tag{1}$$

where x is one of the normalized value in the F_0 information, and y is a transformed value, which is used in the model training. **Figure 5** shows the part of the function (y range is $[-3, 3]$). In the synthesis stage, its inverse function is used for the transformation from the value generated by HTS to target value for speech synthesis. By the way, it is necessary to study the meaning of such non-linear conversion for speech parameters, and if such pre-processing is appropriate for the analysis in HTS. Not only the transformed value, but also the original value has not been the target for such consideration. For example, the agreement between the Euclid distance of a pair of F_0 information and the auditory perceptual distance in F_0 in these segments has not been studied deeply yet. We have tried to find the optimal distance measure by changing it to another one (correlation between the F_0 information [11]), but clear improvement of synthesized speech quality was not confirmed.

It is known that the speech synthesized from the target feature extracted from recorded speech has a "vibration" sounding problem which appears at /no o/ of "hiQshino omoi" in sample J21. (This phenomenon also appears in the speech synthesized from the HTS generated features, around 2.7 s in Figure 4.) These are possible reasons: the low resolution of

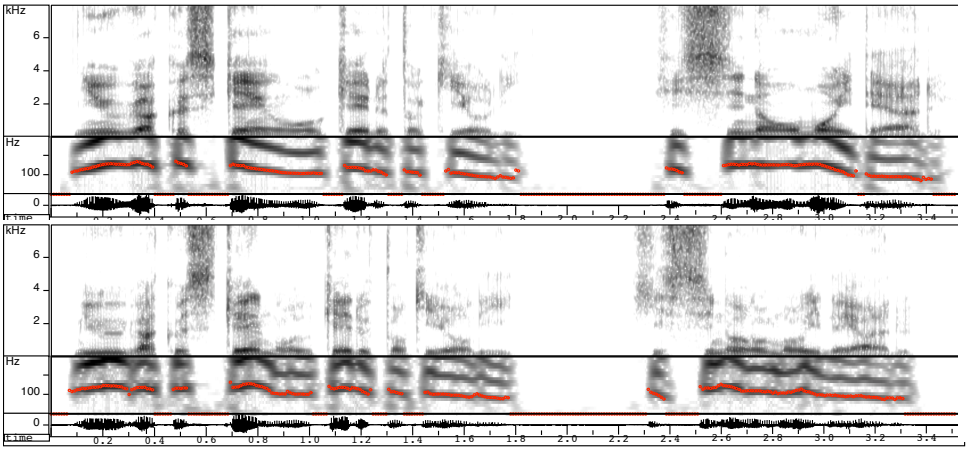


Figure 4: *Synthesized speech (bottom) with original speech (top).*
The utterance is J21: "nyu:gakushIkeNo ukeru tokiyori hiQshino omoide aru (I made more desperate efforts than an entrance examination)." In each panel, the spectrum, enlarged part of narrow band spectrum (0–300 Hz) for F_0 pattern displaying with automatically detected scalar F_0, and waveform are drawn from top to bottom. The unit of time axis is second.

the F_0 information to represent F_0; the limitation of contextual information (currently, only the adjacent segment's distances are considered). It is necessary to investigate if the resolution precision of F_0 affects the speech quality seriously by changing the order of FFT in F_0 analysis.

In the proposed method, the shortest path that exhibits the lowest mean concatenation distortion is searched for exhaustively in the N-best candidates. Such a full search is effective if the distortion measure corresponds to the perceptual measure very well. However, the measure used in this method might not have such correspondence. For this reason, some kind of pruning in the segment search would be effective for improving the naturalness of synthesized speech.

6. Summary

In this paper, we presented a concatenative speech synthesis system in which 5 ms segments are used and an HMM-based feature generation function of HTS is introduced as an LPI-to-feature transformation module. It was confirmed that the reproducibility of the speaker's voice quality was better than that generated by the original HTS.

Since speech synthesized from the extracted features of recorded speech has a vibration sounding problem, the priority for the solution of it should be higher than that of speed-enhancement for synthesis.

7. Acknowledgments

The authors greatly appreciate the discussion with Professor Takayuki Arai (Sophia University, Japan) and Dr. Tomoki Toda (Nara Institute of Science and Technology).

8. References

[1] T. Hirai and S. Tenpaku. Using 5 ms segments in concatenative speech synthesis. In *Proc. 5th ISCA Speech Syntheis Workshop*, June 2004. http://www.ssw5.org/papers/1032.pdf.

[2] Z. Ling and R. Wang. HMM-Based unit selection using frame sized speech segments. In *Proc. ICSLP*, 2006.

[3] T. Hirai. Optimization of target cost weights in concatenative speech synthesis with very short segments of 5-ms duration. In *Proc. 4rd Joint Meeting of ASA and ASJ*, Nov. 2006. Abstract: JASA, Vol. 120, No. 5, Pt. 2, p.3037, 1pSC7.

[4] Z. Ling and R. Wang. HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion. In *Proc. ICASSP*, 2007.

[5] HTS version 2.0, 2006. http:// hts.ics.nitech.ac.jp/.

[6] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuwabara. A large-scale Japanese speech database. In *Proc. ICSLP*, pages 1089–1092, 1990.

[7] K. Tokuda. Reference manual for Speech Signal Processing Toolkit ver. 3.0, 2002. http://kt-labics.nitech.ac.jp/~tokuda/SPTK/.

[8] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, May 2000.

[9] Y. Stylianou and A. K. Syrdal. Perceptual and objective detection of discontinuities in concatenative speech synthesis. In *Proc. ICASSP*, May 2001.

[10] A. V. Aho, J. E. Hopcroft, and J. Ullman. *Data Structures and Algorithms*. Addison-Wesley Pub. Co., 1982.

[11] T. Hirai and S. Tenpaku. Refinement of F0 distance measure in 5 ms segment concatenative speech synthesis. In *Rec. Spring Meeting, Acoust. Soc. Jpn.*, March 2007.