Incremental TTS for Japanese Language

Tomoya Yanagita[1], *Sakriani Sakti*[1,2], *Satoshi Nakamura*[1,2]

[1]Graduate School of Information Science Nara Institute of Science and Technology, Japan
[2] RIKEN, Center for Advanced Intelligence Project AIP, Japan

{yanagita.tomoya.yo8,ssakti,s-nakamura}@is.naist.jp

Abstract

Simultaneous lecture translation requires speech to be translated in real time before the speaker has spoken an entire sentence since a long delay will create difficulties for the listeners trying to follow the lecture. The challenge is to construct a full-fledged system with speech recognition, machine translation, and text-to-speech synthesis (TTS) components that could produce high-quality speech translations on the fly. Specifically for a TTS, this poses problems as a conventional framework commonly requires the language-dependent contextual linguistics of a full sentence to produce a natural-sounding speech waveform. Several studies have proposed ways for an incremental TTS (ITTS), in which it can estimate the target prosody from only partial knowledge of the sentence. However, most investigations are being done only in French, English, and German. French is a syllable-timed language and the others are stress-timed languages. The Japanese language, which is a mora-timed language, has not been investigated so far. In this paper, we evaluate the quality of Japanese synthesized speech based on various linguistic and temporal incremental units. Experimental results reveal that an accent phrase incremental unit (a group of moras) is essential for a Japanese ITTS as a trade-off between quality and synthesis units.

Index Terms: Incremental speech synthesis, linguistic and temporal locality features, HMM based speech synthesis

1. INTRODUCTION

In recent years, the number of international students at universities in Japan has been increasing. But, many lectures are still provided in Japanese, making it difficult for those who do not have a high-level of Japanese language skills to follow the content. One way is to construct an automatic speech-to-speech translation system, which consists of three components: automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) synthesis. ASR usually starts after the speaker has spoken the entire sentence, and MT and TTS perform translation and synthesis sentence by sentence in a consecutive manner [1]. However, spoken speech in a lecture can be very long, thus this method can cause significant delays and a mismatch between descriptions from lecture slides and those of the resulting translation. In contrast, human interpreters generally break sentences into smaller chunks, resulting in a shorter delay. As a solution to this problem, several studies aimed to construct simultaneous speech translation system [2, 3, 4, 5, 6] that could produce high-quality speech translations while at the same time minimizing the latency of the translation process. Consequently, it is necessary to develop real-time ASR, MT and TTS systems. This paper focuses on the challenge of developing a real-time TTS system.

To produce high-quality speech synthesis, many contextual linguistic factors (e.g., phoneme identity and word stress) need to be considered because such information can have an effect on the prosodic characteristics of speech. Specifically, in statistical parametric speech synthesis based on a HMM, the following two processes are typically executed: (a) analyzing the input sentence and extracting linguistic features using natural language processing, and; (b) establishing a sentence-based HMM sequence on the basis of linguistic specifications and estimating acoustic features to generate a speech waveform while performing global optimization so that acoustic features could be changed smoothly [7, 8]. Despite its ability to produce high-quality speech, a conventional TTS system is only able to synthesize speech sentence by sentence because it requires language-dependent contextual linguistic information of a complete sentence and parameter smoothing.

An incremental TTS(ITTS) system, on the other hand, attempts to produce a natural-sounding speech waveform 'on the fly' before receiving a complete sentence. The main challenge is the online estimation of the target prosody from partial knowledge of the sentence's syntactic structure. Specifically, in contrast to the process in a conventional TTS described above, for part (a) the ITTS system has to extract linguistic features in a situation where some (next part-of-speech (POS) tag, the next word, etc.) are unknown during the synthesis. Furthermore, in the processing of (b), a limited HMM sequence has to be constructed from a limited number of linguistic features, local optimization has to be performed, and acoustic features must then be estimated. Unfortunately, the quality of speech may deteriorate due to the limited number of linguistic features and local optimization.

Various studies on ITTS system have proposed possible ways to address the above problems. Baumann et al. investigated how the existence of unknown linguistic features influences the prosodic estimation in English and German ITTS system [9, 10]. Pauget et al. proposed an ITTS training strategy based on HMM with unknown linguistic features [11]. In addition, they also proposed an approach to predict the POS of the next word and use it as a linguistic feature [12]. By adopting the above approaches, ITTS quality has been improved. However, most investigations into the ITTS framework focus only on German, English and French. English and German are stress-timed languages, and French is a syllable-timed language [13]. This study focuses on the Japanese language, which has different prosodic characteristics to German, English, and French.

Japanese is mora-timed and pitch accent language [13, 14]. A mora is a sub-unit syllable consisting of one short vowel and any preceding onset consonant, which corresponds to a single Japanese Hiragana character [15]. An "accent phrase" is a group of moras that is longer than a word unit. The accent of each mora indicates the relative pitch change in the accent phrase and has an essential role in the prosody similar to the tone type in tonal languages[16]. To the best of our knowledge, an ITTS system for mora-timed languages with tonal accent, such as Japanese, does not exist yet. This paper investigates

the effect on speech quality in linguistic and temporal locality choices for a Japanese ITTS system. To have the same framework with most existing ITTS systems, we will also construct our ITTS system with statistical parametric speech synthesis based on HMMs.

2. LINGUISTIC AND TEMPORAL LOCALITY IN JAPANESE ITTS SYSTEM

Shuji et al. states, "In Japanese speech, there is a phonological sound unit named mora, and an accent phrase is determined against a group of several moras. The Japanese accent of each mora indicates the relative pitch change in the accent phrase, and has an important role in the prosody, which is similar to the tone type in tonal languages" [16]. Thus, linguistic labels related to prosody have to be designed on the basis of accent phrase. In a conventional Japanese TTS system, linguistic features usually consist of a phoneme, accent phrase, and breath group (an intonational phrase consisting of several accent phrases). In word-level features, only POS tag information is added as additional information. To realize a Japanese ITTS system, we first define the possible linguistic features in Japanese ITTS and then analyze some possibilities in linguistic and temporal locality choices, both with and without the accent phrase feature.

2.1. Linguistic Information and Process Scope

Figure 1 shows the different scopes of contextual linguistic features that are used to generate TTS and ITTS system. The solid line box in the figure shows a conventional TTS system, given a complete text sentence and its linguistic information. In contrast, an ITTS system aims to generate acoustic features for a speech waveform given only a partial text sentence (Fig. 1, dashed line box).

The details of the contextual linguistic information used in TTS are described below.

Phoneme: Identity of {past, current, next} phoneme.

Word POS: {past, current, next} POS tag information of the word.

Accent phrase: The number of mora in {past, current, next} accent phrase, {past, current, next} accent type of the accent phrase, pause insertion occurrence between accent phrases, {forward, backward} position of mora in the current accent phrase.

Breath group: The number of mora, accent phrases in the {past, current, next} breath group, and the {forward, backward} position of the accent phrase in the {past, current, next} breath group. Note that the breath group is composed of different accent phrases and often divided at the punctuation marks.

Sentence: The number of {mora, accent phrases, breath groups} in the sentence, and {forward, backward} position of breath in the sentence.

The contextual linguistic features that can be used in ITTS are limited as described below.

Phoneme: Identity of {past, current} phoneme.

Word POS: {past, current} POS tag information of the word.

Accent phrase: The number of mora in {past, current} accent phrase, {past, current} accent type of the accent phrase, and {forward, backward} position of mora in the current accent phrase.

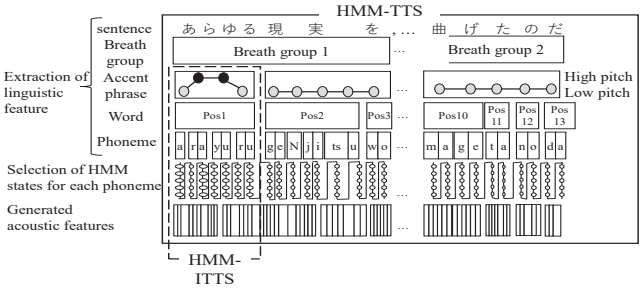


Figure 1: *Different scopes of contextual linguistic features used to generate TTS and ITTS systems.*

Breath group: The number of {mora, accent phrases} in the {past, current} breath group, and the {forward} position of the accent phrase in the {past, current} breath group.

The main difference is the absence of the next and backward linguistic features such as "next POS tag information of the word" or "backward position of breath in the sentence". Furthermore, since we aim to synthesize speech in real time, we do not use any linguistic features that are difficult to detect within the current time-step such as "pause insertion occurrence between accent phrases."

2.2. Linguistic and Temporal Locality

For an ITTS system, the more features we use, the better the quality can be. However, this approach becomes less incremental and the delay gets longer. Therefore, it is important to investigate the optimum linguistic and temporal locality.

First, we classify several possible linguistic locality choices to define the level of linguistic abstraction or the granularity of the "current" chunk. These choices are as follows:

Pho: Only phoneme features.

Pho+POS: Phoneme and word POS.

Pho+Accphr: Phoneme and accent phrase.

Pho+Bre: Phoneme and breath group.

Pho+POS+Accphr: Phoneme, word POS, and accent phrase.

Pho+POS+Bre: Phoneme, word POS, and breath group.

Pho+Accphr+Bre: Phoneme, accent phrase, and breath group.

Pho+POS+Accphr+Bre: Phoneme, word POS, accent phrase, and breath group.

Next, we change the chunks for the ITTS according to the above experimental result, also investigate its speech quality via objective evaluation, and visualize acoustic features in each chunk. Furthermore, to maintain the smoothness between chunks, we also investigate the following possible ways of connecting several chunks while synthesizing text 'on-the-fly' (Figure 2). This includes:

ChunkCurrAccphr: Synthesizing the current chunk of an accent phrase (Figure 2 (a), "Accphr" means each accent phrase).

ChunkCurrAccphr+wPastAccphr: Synthesizing the current chunk of an accent phrase by connecting one chunk of a previous accent phrase (Figure 2 (b)).

ChunkCurrAccphr+wAllPastAccphr: Synthesizing the current chunk of an accent phrase by connecting all chunks from previous accent phrases (Figure 2 (c)).

ChunkCurrAccphr+wPastAccphr+wNextAccphr: Synthesizing the current chunk of an accent phrase by connecting one chunk of a previous accent phrase and one of a next accent phrase (Figure 2 (d)).

3. EXPERIMENTAL SETUP

We used the ATR 503 phonetically balanced sentences [17] of the HTS demo [18] as the dataset, from which 450 sentences were used as the training set while the rest were used for the test set. The speech features include 39-dimensional mel-cepstrum coefficients, 1-dimensional fundamental frequency, 5-dimensional aperiodic components, and their respective dynamic features. We also used STRAIGHT [19] to extract speech features, and used the HTS engine for speech synthesis.

First, we synthesized speech using a standard TTS system, and used the results as a reference. After that, we synthesized speech using various ITTS systems. A perceptual-based measure of the differences in terms of fundamental frequency (F0) between TTS and ITTS systems was calculated as follows:

$$C_{f_0} = \frac{1}{T}\sum_{t-1}^{T} 1200 log_2 \frac{|f_0^{tar}(t)|}{|f_0^{src}(t)|}, \quad (1)$$

where F0 is evaluated in cent between two speeches, and 1200 cents represents the difference of 1 octave[11]. Furthermore, we also calculated the accuracy of the estimated spectrum using a mel-cepstral distortion[20] in dB which is defined as follows:

$$MCD = \frac{1}{T}\frac{10}{ln(10)}\sum_{t=1}^{T}\sqrt{2\sum_{d=0}^{D}(y_{t,d}^{tar} - y_{t,d}^{src})^2}, \quad (2)$$

where t and d are the number of frames and mel-cepstrum dimensions, respectively. In addition to an objective evaluation, we also conducted a mean opinion score (MOS) test as a subjective evaluation [21]. The subjective evaluation was conducted with 16 native Japanese speakers. In total, there were 45-60 speech utterances (15 utterances per system), which were presented in random order. Subjects listened to each presented speech and were required to rate the overall quality with regard to naturalness. A 5-point MOS scale was used, where 5 indicated excellent (the speech utterance sounds very clear and perfectly natural) and 1 indicated bad (the speech utterance sounds unclear and completely unnatural). Each speech utterance could be played as many times as the subjects wished.

4. Objective and Subjective Assessments of Synthesized Speech

4.1. Investigation of Linguistic Locality

We investigated the classes we defined in Section 2.2 that represent the level of linguistic abstraction or the granularity of the "current" chunk. The objective and subjective evaluations were done on the basis of a comparison with the standard TTS model using full-context linguistic information. The global optimization by dynamic features was affected in these evaluations because the "current" chunks are sentences. In this experiment, generated acoustic features were considered to be the

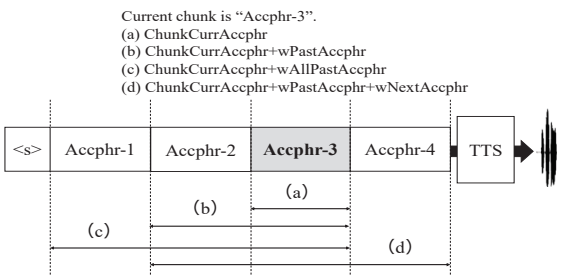


Figure 2: *Possible ways of connecting several chunks while synthesizing a text 'on-the-fly'.*

Table 1: *Objective and subjective evaluation for ITTS systems with various linguistic context abstraction.*

Linguistic feature	C_{f_0}[cent]	MCD[dB]	MOS
Pho	242.5	3.5	-
Pho+POS	211.2	3.5	2.2
Pho+Accphr	178.8	3.4	-
Pho+Bre	186.8	3.5	
Pho+POS+Accphr	141.1	3.4	3.2
Pho+POS+Bre	175.3	3.4	-
Pho+Accphr+Bre	83.9	3.3	3.4
Pho+POS+Accphr+Bre	84.2	3.3	-
General TTS	0.0	0.0	3.6

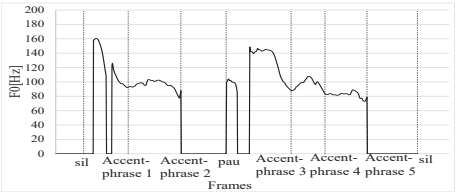
upper limit speech quality of a Japanese ITTS system. In other words, we assume it has a lower speech quality than that of a conventional TTS.

Table 1 shows the results of F0 difference average and MCD average in the objective evaluation and MOS average in the subjective evaluation. The results show that there is not really much difference when we use only phoneme or phoneme with word POS as the level of linguistic abstraction (see "Pho" vs "Pho+POS"). However, upgrading the information to accent phrase level could improve the prosody quality (see "Pho+POS+Accphr"). As expected, the best performance was the one using the longest context information (see "Pho+Accphr+Bre"). Nevertheless the MCD value of those systems are almost the same.

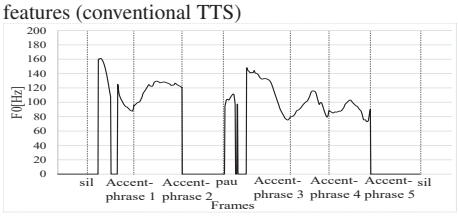
On the basis of the results of Table 1, we selected three variations of contextual linguistic features ("Pho+POS", "Pho+POS+Accphr", and "Pho+Accphr+Bre") and performed a subjective evaluation. For comparison, we also included the MOS test for the standard TTS model with full-context linguistic information. The result of the ITTS system with only "Pho+POS" linguistic information was lowest. The results revealed that it is difficult to construct a Japanese ITTS system that produces synthesized speech word by word. The MOS scores of ITTS with "Pho+POS+Accphr" versus "Pho+Accphr+Bre" linguistic information were quite close. This indicates that the minimum level of linguistic abstraction would be in the accent phrase. Surprisingly, the MOS scores of "Pho+POS+Accphr" and "Pho+Accphr+Bre" are not that different from that of the standard TTS, they do not have any information of the next linguistic features. However, this might be the effect of the global optimization by dynamic features. We will thus choose the accent phrase as the synthesis unit, and further investigate its effect in the next experiment.

4.2. Investigation of Temporal Locality

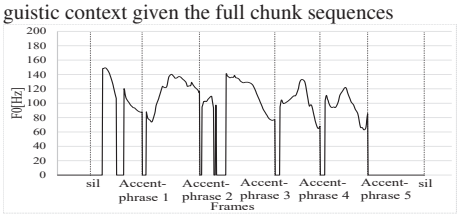
Next, we focused on an ITTS system based on the accent-phrase unit. We investigated a speech quality of Japanese ITTS, temporal locality choices, to define an accent phrase as the "current" chunk. In this case, ITTS uses local optimization instead of global optimization for parameter estimation. Since we synthesized text accent phrase by accent phrase, the occurrence of a prosody break between accent phrases would become possible. After that, we connected the chunk-based synthesized speech to synthesized sentence speech. The first row of Table 2 shows the averages for F0 difference, MCD, and MOS for ITTS with only an accent phrase as the "current" chunk. As we expected, the F0 difference and MCD values are worse than those for "Pho+POS+Accphr" in Table 1. Figure 3 shows a comparison of the generated F0 sequences based on different linguistic and temporal locality: (a) F0 sequences based on full contextual



(a) F0 sequences based on full contextual linguistic features (conventional TTS)



(b) F0 sequences based on "Pho+POS+Accphr" linguistic context given the full chunk sequences



(c) F0 sequences based on "Pho+POS+Accphr" linguistic context and current accent phrase chunk ("ChunkCurrAccphr") only

Figure 3: *Comparison of generated F0 sequences based on different linguistic and temporal locality.*

linguistic features (conventional TTS); (b) F0 sequences based on "Pho+POS+Accphr" linguistic context given the full chunk sequences; (c) F0 sequences based on "Pho+POS+Accphr" linguistic context and current accent phrase chunk ("ChunkCurrAccphr"). Here, "sil", "pau", and "accent phrase" indicate frame boundary of silence, pause, and accent phrase, respectively. The results of Figure 3 indicate that smooth f0 sequences can be generated using a longer temporal locality (see Figure 3(a) and Figure 3(b)). However, a prosody break occurs when using only the chunk of current accent phrase units (see Figure 3(c)). This suggests that we need to find a way to maintain smoothness between the chunks, which will be discussed in the next session.

4.3. Investigation of Chunk Connection

In the previous section, we used "Pho+POS+Accphr" as contextual linguistic information. We then investigate possible ways of connecting several chunks while synthesizing text 'on the fly' (Figure 2). Table 2 shows the F0 difference and MCD for ITTS systems with different ways of connecting the chunks. The results reveal that "ChunkCurrAccphr+wPastAccphr+wNextAccphr" gave the best performance. By considering the past and next accent phrase units, a smoother prosodic estimate could be made among the accent phrases.

The results of the subjective evaluation show a similar tendency to those of the objective evaluation. Connecting all past accent phrase chunks could improve naturalness. Nevertheless, the best system is achieved with "ChunkCurrAccphr+wPastAccphr+wNextAccphr." This suggests the need to wait one chunk before starting the synthesis process. Figure 4 shows f0 sequences of "ChunkCurrAccphr+wPastAccphr+wNextAccphr." As we expected, f0 sequences are also much smoother than those for an accent phrase ITTS (Figure 3(c)).

Table 2: *Objective and subjective evaluation of ITTS systems with various chunk connections.*

Chunk connection	C_{f_0}[cent]	MCD[dB]	MOS
ChunkCurrAccphr	232.6	5.2	2.7
ChunkCurrAccphr +wPastAccphr	170.5	4.5	-
ChunkCurrAccphr +wAllPastAccphr	160.8	4.2	2.8
ChunkCurrAccphr +wPastAccphr +wNextAccphr	157.3	4.0	3.3

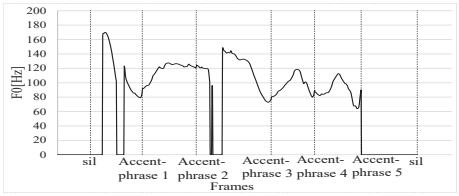


Figure 4: *Smoother f0 sequences by chunk connection.*

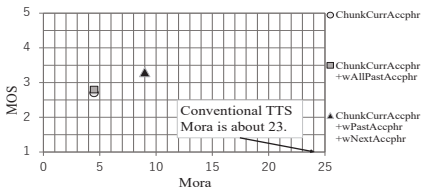


Figure 5: *Trade-off between delay and speech quality for ITTS system with various chunk connections.*

4.4. Investigation of Chunk Connection vs Time Delay

Finally, we investigated the impact of time delay on our Japanese ITTS system given the limited number of contextual features. To clarify the differences, we used the number of mora as the units. Figure 5 shows the average mora per accent phrase in the test set and the Mos score at the chunk connection. The number of mora per sentence in the test set is about 23. The chunk connection could help to improve the speech quality while having a smaller number of mora than that in the conventional TTS.

5. CONCLUSIONS

This paper presents the first investigation of an ITTS system for mora-timed languages such as Japanese. Our research aimed to find the optimum strategy for developing a Japanese ITTS system when given non-complete contextual linguistic features. We explored various ways of using limited contextual linguistic and temporal features. The experimental results revealed that using a word-by-word synthesizer for the Japanese language does not provide good-quality speech. We found that the linguistic feature of accent phrase is critical when the next linguistic features are missing, since the accent phrase is required for estimating prosody. Therefore, we defined our unit chunk as being on the basis of the accent phrase. Furthermore, since the past and next accent phrase units are considered, a smoother prosodic estimate could be made among the accent phrases; consequently, this study suggested the need to wait one more chunk before starting the synthesis process. In the future, we will further investigate the possibility of developing a Japanese ITTS system using a deep learning framework.

6. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP 17K00237.

7. References

[1] E. Matusov, A. Mauser, and H. Ney, "Automatic sentence segmentation and punctuation prediction for spoken language translation," in *in Proc. IWSLT*, 2006, pp. 158–165.

[2] C. Fügen, A. Waibel, and M. Kolss, "Simultaneous translation of lectures and speeches," *Machine Translation*, vol. 21, no. 4, pp. 209–252, Dec 2007.

[3] S. Bangalore, V. K. Rangarajan Sridhar, P. Kolan, L. Golipour, and A. Jimenez, "Real-time incremental speech-to-speech translation of dialogs," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ser. NAACL HLT '12, Stroudsburg, PA, USA, 2012, pp. 437–445.

[4] T. Fujita, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "Simple, lexicalized choice of translation timing for simultaneous speech translation," in *14th Annual Conference of the International Speech Communication Association (InterSpeech 2013)*, Lyon, France, August 2013, pp. 3487–3491. [Online]. Available: http://www.phontron.com/paper/fujita13interspeech.pdf

[5] H. Shimizu, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "Constructing a speech translation system using simultaneous interpretation data," in *10th International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, December 2013, pp. 212–218. [Online]. Available: http://www.phontron.com/paper/shimizu13iwslt.pdf

[6] Y. Oda, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "Optimizing segmentation strategies for simultaneous speech translation," in *The 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, USA, June 2014, pp. 551–556. [Online]. Available: http://www.phontron.com/paper/oda14acl.pdf

[7] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1315–1318.

[8] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[9] T. Baumann, "Partial representations improve the prosody of incremental speech synthesis," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[10] ——, "Decision tree usage for incremental parametric speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3819–3823.

[11] M. Pouget, T. Hueber, G. Bailly, and T. Baumann, "HMM training strategy for incremental speech synthesis," in *16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, 2015, pp. 1201–1205.

[12] M. Pouget, O. Nahorna, T. Hueber, and G. Bailly, "Adaptive latency for part-of-speech tagging in incremental text-to-speech synthesis," in *Interspeech 2016*, 2016, pp. 2846–2850.

[13] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology*, vol. 7, no. 515-546, 2002.

[14] D. Hirst and A. Di Cristo, *Intonation systems: a survey of twenty languages*. Cambridge University Press, 1998.

[15] M. Suzuki, R. Kuroiwa, K. Innami, S. Kobayashi, S. Shimizu, N. Minematsu, and K. Hirose, "Accent sandhi estimation of tokyo dialect of japanese using conditional random fields," *IEICE TRANSACTIONS on Information and Systems*, vol. 100, no. 4, pp. 655–661, 2017.

[16] S. Yokomizo, T. Nose, and T. Kobayashi, "Evaluation of prosodic contextual factors for HMM-based speech synthesis," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[17] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[18] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (hts) version 2.0." in *SSW*, 2007, pp. 294–299.

[19] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[20] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Communications, Computers and Signal Processing, 1993., IEEE Pacific Rim Conference on*, vol. 1. IEEE, 1993, pp. 125–128.

[21] I. Recommendation, "800, methods for subjective determination of transmission quality," *International Telecommunication Union*, 1996.