IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 February 2023, accepted 21 February 2023, date of publication 2 March 2023, date of current version 9 March 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3251657*

RESEARCH ARTICLE

# Japanese Neural Incremental Text-to-Speech Synthesis Framework With an Accent Phrase Input

**TOMOYA YANAGITA[1], SAKRIANI SAKTI[2], (Member, IEEE), AND SATOSHI NAKAMURA[1], (Fellow, IEEE)**

[1]Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan
[2]Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan

Corresponding author: Sakriani Sakti (ssakti@jaist.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP21H05054 and Grant JP21H03467.

**ABSTRACT** Work in the development of neural incremental text-to-speech (iTTS), which is attracting increasing attention, has recently pursued low-latency processing by generating speech on the fly before reading complete sentences. Most current state-of-the-art iTTS systems use a prefix-to-prefix neural iTTS framework with look-ahead of 1-2 unit segments (i.e., phonemes or words). However, since the Japanese language is based on accent phrase units that are longer than words, using a prefix-to-prefix neural iTTS with a look-ahead approach increases latency. Here, we propose an alternative to the end-to-end neural iTTS architecture that does not apply look-ahead input when synthesizing speech chunks. We further propose a method to use information from the previous time step by connecting the synthesized vector and the model's internal state to the current time step. We experimentally investigated the latency of various iTTS systems with different modeling and synthesis chunks. The experimental results show that, for Japanese, the proposed iTTS is able to synthesize better speech quality, with a similar latency range, than the conventional baseline prefix-to-prefix neural iTTS with word units. Moreover, we found that our proposed approach improved the prosodic naturalness among synthesized units in the Japanese language. Subjective evaluations also revealed that the proposed approach with an incremental unit of two accent phrases achieved the best scores in Japanese iTTS systems.

**INDEX TERMS** Incremental speech synthesis, end-to-end, Japanese language, accent phrase unit.

## I. INTRODUCTION

Speech-to-speech translation (S2ST) is an innovative technology that translates speech signals from a source language to another language, enabling people of different languages to communicate in their native tongues. S2ST systems commonly consist of three components [1]: automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) synthesis. In operation, such a system first recognizes the source language speech as a source language text, automatically translates this into a target language text, and finally synthesizes the target language speech by TTS. In conventional systems, the process is done sentence by sentence. Conventional S2STs produce translated speech with significant latency for the longer speech segments of lectures and meetings, thus creating difficulty for listeners who are struggling to follow the speaker's talk or an ongoing conversation.

In contrast to conventional S2ST systems, human interpreters generally break sentences into smaller chunks and incrementally translate them based on partial information with minimum latency [2]. Researchers have recently started to develop incremental speech-to-speech translation/interpretation systems toward a simultaneous

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

interpretation process for humans. One critical difference from standard S2ST systems is that, in an incremental approach, each component (ASR, MT, TTS) needs to generate on-the-fly output before it receives a complete sentence. In this paper, we focus on incremental TTS (iTTS).

End-to-end TTS systems [3], [4], [5], [6], [7] have been proposed based on sequence-to-sequence modeling. Unlike parametric TTS systems [8], [9], [10], the neural end-to-end architecture simplifies models so that the neural network directly maps input features to speech outputs or acoustic feature outputs. Developments of neural vocoders [11], [12], [13], [14], [15], [16] to reconstruct speech from acoustic features or a noise sequence have also made remarkable progress. Therefore, the speech quality of the end-to-end architecture has reached the level of human speech [4].

Hidden Markov models (HMMs) have previously been applied to iTTS [17], [18], [19], [20]. Taking into account the performance improvement by end-to-end TTS systems, sequence-to-sequence modeling has also been applied to iTTS recently [21], [22], [23], [24], [25], [26]. In English, a prefix-to-prefix framework [22] was proposed that allows waits for look-ahead of 1-2 words in iTTS. Although this prefix-to-prefix iTTS used phoneme sequences as input and produced good speech quality, it could not automatically control the look-ahead length. Another work [24] proposed a prefix-to-prefix iTTS with reinforcement learning to control the tradeoff between look-ahead words and speech quality. Other research [23] analyzed the look-ahead effects in a prefix-to-prefix iTTS and found that the look-ahead word length significantly affected quality. From this analysis, Stephenson et al. [25] proposed a method that predicts look-ahead text using a language model. A similar method [26] was then also proposed. These related works [22], [23], [24], [25], [26] use phoneme sequences as input features and word units for a synthesis chunk.

In a Japanese iTTS, based on the HMM framework [20], the input features are sequences of phonemes and the linguistic features of accent phrase units. Since accent phrases are longer than a word, a word-based, prefix-to-prefix neural network cannot be simply applied to a Japanese iTTS. When we apply the prefix-to-prefix neural network to Japanese, the look-ahead length is 1-2 accent phrases. Using 1-2 accent phrases with the look-ahead approach does not produce an unacceptable latency. We previously presented a preliminary result of a neural Japanese iTTS system [21] that uses accent phrases and phonemes without a look-ahead approach. This paper is an extended version with newer modeling, a deeper analysis of the synthesis unit, and comparisons of latency and quality with those of the related works. Furthermore, we propose an additional method, using various accent features in addition to accent phrases, and use Parallel WaveGan [15] and Tacotron2 [4] to improve speech quality. For the Japanese baseline prefix-to-prefix iTTS, we use a morpheme unit as the synthesis chunk to minimize latency and then compare it to our proposed approaches. The latency and quality of the proposed iTTS are analyzed and then compared to the results of prefix-to-prefix iTTS.

## II. JAPANESE iTTS INPUTS

Japanese is a pitch-accent aspect language. Its accent of each mora, which indicates the relative pitch change in the accent phrase, plays an important role in prosody, which resembles tone types in tonal languages [27]. One mora is approximately equivalent to one hiragana character. The Japanese word ''ha shi'' can mean either a bridge or a pair of chopsticks. If the pitch changes from high to low, it means a bridge; if the pitch changes from low to high, it means a pair of chopsticks. Such pitch information is represented not in words or phonemes but in accent phrase units, which have one accent type that changes a pitch from high to low. An accent phrase has only one accent type, and the type depends on the context. Accent type is critical for representing the meaning in a given context, although it does not represent a phoneme sequence. A Japanese TTS needs another input for the accent phrase.

Fig. 1 shows a converting process from a surface text to inputs: a phoneme and an accent type in the accent phrase. The Japanese surface text has no word boundary, such as a space character in English, and a morpheme unit is useful to obtain the phoneme and the accent type. Consequently, morpheme analysis detects morphemes in the text, and each morpheme includes pronunciation, a part-of-speech tag, and accent information for constructing the accent type. To obtain features for an accent phrase, morpheme units are reconstructed by mora units. Pronunciations are converted to a sequence of phonemes with a pronunciation dictionary. Furthermore, the accent phrase boundary is detected by the part-of-speech tags, and the accent type of the accent phrase is obtained from rules with part-of-speech and accent information of the mora unit. Finally, phonemes and accent types are used as inputs for the Japanese TTS, and using features in the accent phrase for Japanese end-to-end TTS systems is known to improve speech quality [28], [29], [30].

Features of the accent phrase are assigned to each phoneme and defined on the mora unit. Therefore, the same feature of the accent phrase will be assigned to each phoneme. Fig. 2 shows accent phrase features of ''kyo o wa'' (''Today is'') on mora units in the accent phrase. We use five features in the accent phrases. A1 is the difference between the position of the phoneme in mora units and the positions of the accent type. For example, A1 of the phoneme ''o'' in the second mora unit is 1 because the difference between the position of the phoneme ''o'' in the mora unit and the mora's position of the accent type is $2 - 1 = 1$. We expect A1 to increase the number of features related to the accent type and mora units. A2 and A3 are the forward and backward positions of the mora in the accent phrase, that is, A2 of the phoneme ''o'' in the second mora unit is 2, and A3 of the phoneme ''o'' in the second mora unit is also 2. A4 is the number of moras in the accent phrase, and A5 is the accent type of the accent phrase.

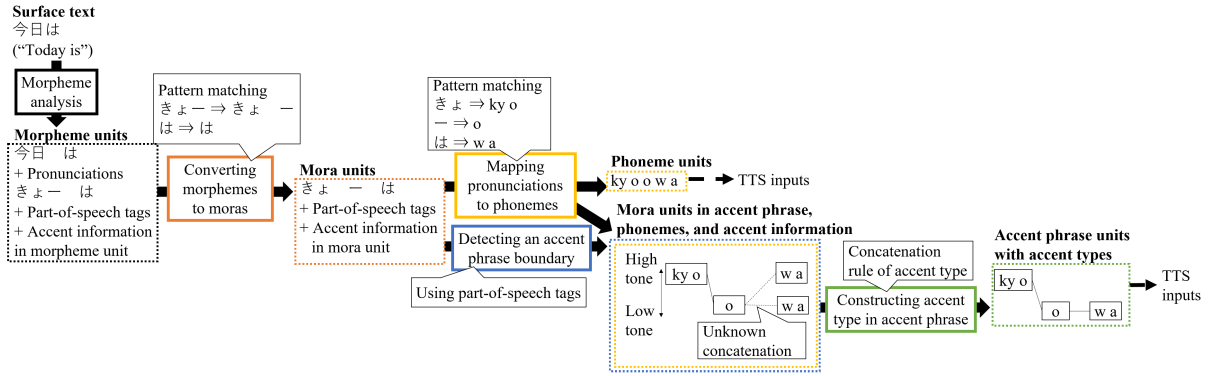



**FIGURE 1.** Flowchart of extracting Japanese TTS inputs.

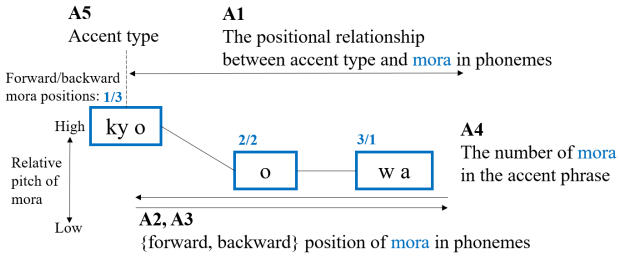



**FIGURE 2.** Japanese accent features in an accent phrase.

## III. METHODS

This paper deals with neural TTS, which consists of two steps: a neural encoder-decoder model to infer acoustic features from input sequences and a neural vocoder to synthesize speech from acoustic features.

The baseline prefix-to-prefix iTTS uses a word unit as the synthesis unit [22]. As described in Section II, a morpheme unit is useful in obtaining inputs for Japanese TTS systems. In later experiments, the Japanese baseline prefix-to-prefix iTTS used morpheme units as the synthesis chunk instead of word units.

### A. SENTENCE-BASED TTS

A sentence-based TTS processes a text sentence by sentence, that is, a synthesis chunk is a full sentence. A full sentence containing $N$ words is represented with a sequence of words $x_{1:N} = [\overline{x_1}, \ldots, \overline{x_N}]$, where the word $\overline{x_t} = [x_t^1, \ldots, x_t^{s_t}]$ includes an input sequence of phonemes with a length $s_t$.

The encoder transforms the input sequence into another feature sequence as hidden states $\mathbf{h}_{1:N} = Enc(x_{1:N}) = [\overline{\mathbf{h}_1}, \ldots, \overline{\mathbf{h}_N}] = [\mathbf{h}_1^1, \ldots, \mathbf{h}_1^{s_1}, \ldots, \mathbf{h}_N^1, \ldots, \mathbf{h}_N^{s_N}]$, where $Enc(\cdot)$ represents the encoder's process.

After getting the encoder's hidden states, the decoder infers acoustic features. The chunk of acoustic features for the word $\overline{\mathbf{y}_t} = [\mathbf{y}_t^1, \mathbf{y}_t^2, \ldots]$ is estimated by $\mathbf{h}_{1:N}$ and $\mathbf{y}_{<t}$, where $\mathbf{y}_{<t} = [\overline{\mathbf{y}_1}, \ldots, \overline{\mathbf{y}_{t-1}}]$ are sequences of acoustic features until the previous words. More specifically, the $i$-th frame for any word is as follows:

$$\mathbf{y}_t^i = Dec(\mathbf{h}_{1:N}, \mathbf{y}_{<t} \circ \mathbf{y}_{t,<i}), \tag{1}$$

where $Dec(\cdot)$ denotes the decoder's process, $\mathbf{y}_{t,<i} = [\mathbf{y}_t^1, \ldots, \mathbf{y}_t^{i-1}]$, and $\circ$ is the concatenation of two sequences. Finally, the sentence's speech waveform $w_{1:N} = [\overline{w_1}, \ldots, \overline{w_N}] = [w_1^1, w_1^2, \ldots, ]$ is as follows:

$$w_{1:N} = \phi(\mathbf{y}_{1:N}), \tag{2}$$

where $\phi(\cdot)$ represents the neural vocoder's process, $\mathbf{y}_{1:N}$ are acoustic features of the full-sentence.

### B. WORD-BASED iTTS

Unlike sentence-based TTS, iTTS uses a partial synthesis chunk instead of the full sentence. We use a prefix-to-prefix iTTS as a baseline, where the synthesis chunk is one word (top table in Fig. 3). The prefix-to-prefix iTTS uses a look-ahead approach to account for the following speech changes. The look-ahead approach waits for $k$ words before the encoder process. The look-ahead length is determined by the following function:

$$g(t) = min\{t + k, |\overline{x_{1:N}}|\}, \tag{3}$$

where $|\overline{x_{1:N}}|$ indicates the total number of words in the sentence.

Under the condition of the look-ahead approach, the sequence of hidden states for the word is represented by $\mathbf{h}_{1:g(t)} = Enc(x_{1:g(t)}) = [\overline{\mathbf{h}_1}, \ldots, \overline{\mathbf{h}_{g(t)}}]$. In other words, the sequence of hidden states is conditioned by the $g(t)$ words. Therefore, the $i$-th acoustic feature for the word and the speech waveform of the word are as follows:

$$\mathbf{y}_t^i = Dec(\mathbf{h}_{1:g(t)}, \mathbf{y}_{<t} \circ \mathbf{y}_{t,<i}), \tag{4}$$

$$\overline{w_t} = \phi(\overline{\mathbf{y}_t}). \tag{5}$$

### C. PROPOSED ACCENT-PHRASE-BASED iTTS

As described earlier, an accent phrase is important for representing Japanese intonation and meaning. Therefore, we propose Japanese iTTS on the basis of the accent phrase as the synthesis chunk. In contrast to subsection III-B, the full

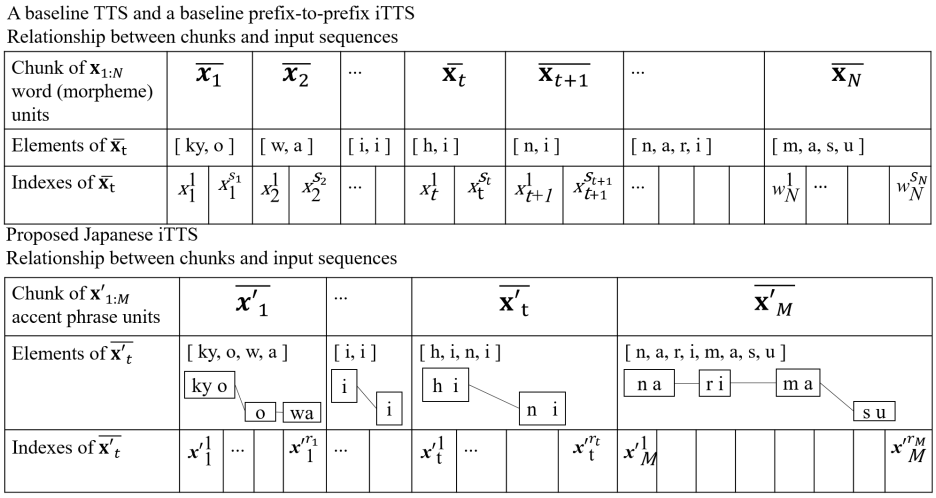



**FIGURE 3.** Synthesis chunks and their elements for a sentence-based TTS, a prefix-to-prefix iTTS, and the proposed accent-phrase-based iTTS.

sentence containing $M$ accent phrases is represented with a sequence of the accent phrase $\mathbf{x}'_{1:M} = [\overline{\mathbf{x}'_1}, \ldots, \overline{\mathbf{x}'_M}]$, where the accent phrase $\overline{\mathbf{x}'_t} = [\mathbf{x}'^1_t, \ldots, \mathbf{x}'^{r_t}_t]$ includes an input sequence of phonemes with length $r_t$ and a sequence of accent features (bottom table in Fig. 3). We propose two methods to estimate acoustic features for the accent phrase.

The first method is **dec+in**. The encoder's hidden states of the accent phrase are observed:

$$\overline{\mathbf{h}'_t} = Enc(\overline{\mathbf{x}'_t}) = [\mathbf{h}'^1_t, \ldots, \mathbf{h}'^{r_t}_t]. \tag{6}$$

Acoustic features in the accent phrase $\overline{\mathbf{y}'_t} = [\mathbf{y}'^1_t, \mathbf{y}'^2_t, \ldots]$ are estimated by $\overline{\mathbf{h}'_t}$ and the last acoustic feature for the previous accent phrase $\mathbf{y}'^p_{t-1}$. The $i$-th acoustic feature for the accent phrase is the following:

$$\mathbf{y}'^i_t = Dec(\overline{\mathbf{h}'_t}, \mathbf{y}'^p_{t-1} \circ \mathbf{y}'_{t,<i}), \tag{7}$$

where $\mathbf{y}'_{t,<i} = [\mathbf{y}'^1_t, \mathbf{y}'^2_t, \ldots, \mathbf{y}'^{i-1}_t]$. The $\overline{\mathbf{h}'_t}$ does not use the hidden vectors in the previous accent phrase $\overline{\mathbf{x}'_{t-1}}$.

Fig. 4 (a) shows how **dec+in** functions in the Japanese iTTS system. The first accent phrase starts from the beginning of the sentence, and other accent phrases start from its middle. We set the initial decoder's inputs as the Mel spectrogram's last frame from the previous accent phrase.

The second method, **dec+in+hidden**, connects not only the last acoustic feature but also the previous states of the model to the current states of the model (Fig. 4 (b)). Therefore, the encoder's hidden states and the acoustic feature are as follows:

$$\overline{\mathbf{h}'_t} = Enc(\mathbf{x}'_{1:t-1} \circ \overline{\mathbf{x}'_t}), \tag{8}$$

$$\mathbf{y}'^i_t = Dec(\overline{\mathbf{h}'_t}, \mathbf{y}'_{<t} \circ \mathbf{y}'_{t,<i}), \tag{9}$$

where $\mathbf{y}'_{<t} = [\overline{\mathbf{y}'_1}, \ldots, \overline{\mathbf{y}'_{t-1}}]$.

$\mathbf{y}'^p_{t-1}$ in (7) and $\mathbf{y}'_{<t}$ in (9) are different due to encoding processes in (6) and (8). In using the second method, we expect it to learn not only the acoustic feature time-series but also the model's internal state change.

Finally, the speech waveform of the accent phrase is the following:

$$\overline{w_t} = \phi(\overline{\mathbf{y}'_t}). \tag{10}$$

## IV. EXPERIMENT

### A. EXPERIMENTAL CONDITIONS

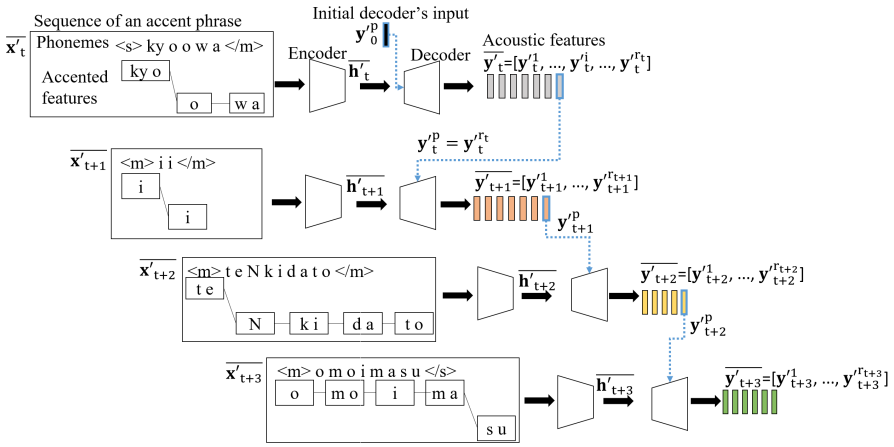
#### 1) DATASET AND MODELS

We used the JSUT dataset (version 1.1), which has 7,696 sentences (10 hours of audio sampled at 48-kHz, which we down-sampled to 22.05-kHz) spoken by a single native female speaker [31]. The data were divided into 7,196 pairs (speech and input sequences) for training, 250 pairs for the development set, and 250 pairs for the test set. We used Open Jtalk[1] for extracting the phoneme and accent features from the text and files with speech duration.[2] We used a Geforce RTX TITAN with a memory of 24 gigabytes.

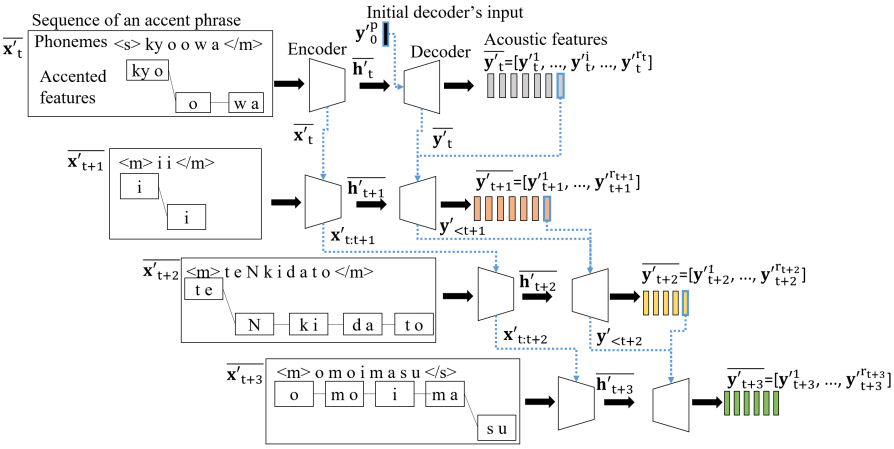
The acoustic features were extracted by Fourier transform, and our final set was composed of 80 dimensions of log Mel spectrogram features. The size of the Fourier transform was 2,048 points. The frameshift and frame lengths were 10 and 50 milliseconds, respectively. We used Tacotron2 [4] to estimate acoustic features from inputs, and a Parallel WaveGan [15] to reconstruct speech from the acoustic features. Unlike the original Tacotron2, we used a uni-directional LSTM to connect hidden states of the model and Forward Attention with Transit Agent [32] to quickly converge the attention. We used an Adam [33] optimizer with a 32-batch size. The learning rate was 1e-3.

We used the prefix-to-prefix model as a baseline iTTS. The input feature is a phoneme sequence (**Pho**).

[1]Open Jtalk – http://open-jtalk.sourceforge.net/
[2]https://github.com/r9y9/jsut-lab




(a) **dec+in**: connecting the last synthesis vector to the decoder's initial input




(b) **dec+in+hidden**: connecting not only the last synthesis vector but also the model's internal states to the encoder and decoder

**FIGURE 4.** **Proposed approaches to Japanese iTTS.**

To accommodate such pitch information, we used two input types of Japanese iTTS to improve speech quality. **Pho+AccType** uses phonemes and only accent types (A5) in accent phrases. We used two embedding layers for the phonemes and the accent types. Then we concatenated two embedding outputs as single input. **Pho+AccFeats** uses phonemes and both accent types (A5) and many accent features (A1, A2, A3, and A4) in the accent phrases. We used six embedding layers for the phonemes and many accent features and concatenated the embedding outputs as one input. Table 1 shows the size of each embedding layer in our experiment. The first column indicates input features, and the second indicates the size of each embedding layer. We replaced low-frequency input features with an unknown symbol using a threshold to deal with unknown inputs. In Japanese iTTS, the vocabulary size is increased by two due to the special characters that indicate the middle.

Baseline TTS and iTTS systems were trained in sentence-based units. The speech was synthesized using each synthesis chunk by adding location symbols to differentiate the unit's location: <s> is the sentence's start and </s> is the sentence's end. The terminating process of the decoder differs in each model. The decoding process in the TTS is controlled by the stop flag [4]. The iTTS uses the stop flag and the alignment distribution to stop the decoder in order to synthesize Mel spectrogram frames [22].

**TABLE 1.** **Input feature types and embedding dimensions.**

| Input feature types | Input and embedding dimensions |
|---|---|
| Pho | Phoneme feature dimension: 44,<br>embedding dimension: 512 |
| Pho+AccType | Phoneme feature dimension: 44 (TTS) or 46 (iTTS),<br>embedding dimension: 480<br>A5-feature dimension: 23, embedding dimension: 32 |
| Pho+AccFeats | Phoneme feature dimension: 44 (TTS) or 46 (iTTS),<br>embedding dimension: 432<br>A1-feature dimension: 26, embedding dimension: 16<br>A2-feature dimension: 20, embedding dimension: 16<br>A3-feature dimension: 20, embedding dimension: 16<br>A4-feature dimension: 23, embedding dimension: 16<br>A5-feature dimension: 23, embedding dimension: 16 |

On the other hand, our proposed method was trained on the accent-phrase-based units, and the decoding process was controlled only by the stop flag. The speech was synthesized using the accent phrase by adding location symbols to differentiate the unit's location: <s>, </s>, <m> is the middle sentence's start, and </m> is the middle sentence's end. When we used **dec+in**, we connected only the Mel spectrogram to each synthesis chunk. When we used **dec+in+hidden**, we connected not only the Mel spectrogram but also the RNN hidden states on each synthesis chunk.

#### 2) EVALUATION INDEXES

We used natural speech as a reference in our objective evaluation of speech quality. Then, we synthesized the speech using various iTTS systems. We calculated a perceptual-based measure in terms of the fundamental frequency (F0) between natural speech and synthesized speech as follows:

$$C_{f_0} = \frac{1}{T}\sum_{t=1}^{T} 1200 log_2 \frac{|f_0^{tar}(t)|}{|f_0^{src}(t)|}, \tag{11}$$

where $f_0^{tar}$ is F0 of natural speech, $f_0^{src}$ is F0 of synthesized speech, and 1200 cents represents a difference of 1-octave [18]. We also calculated the accuracy of the estimated spectrum using Mel cepstrum distortion [34] in dB, defined as follows:

$$MCD = \frac{1}{T}\frac{10}{ln(10)}\sum_{t=1}^{T}\sqrt{2\sum_{d=1}^{D}(y_t^d - \hat{y}_t^d)^2}, \tag{12}$$

where $t$ and $d$ are the number of frames and the Mel cepstrum dimensions, respectively. $y$ is a Mel cepstrum component of natural speech, $\hat{y}$ is a Mel cepstrum component of synthesized speech.

In our subjective evaluation of speech quality, we calculated a mean opinion score (MOS) test [35] for the naturalness of changing lengths of incremental units. Subjects listened to each presented bit of speech audio and rated the overall quality based on its naturalness. We used a 5-point MOS scale, where 5 indicated excellent speech utterances (very clear and completely natural) and 1 indicated bad speech utterances (unclear and completely unnatural). We conducted subjective evaluations in Japanese with 13 native speakers. Synthesized speech samples were evaluated from 10 speech utterances per model.

To analyze latency, we used a re-speaking system that synthesized the same speech after playing natural speech with each chunk. We measured the latency time from starting an input sequence to finishing each bit of synthesis speech. Then we calculated the frequency of latency in each model.

### B. OBJECTIVE EVALUATION OF METHODS

As described above, we used two types of input features in Fig. 2 and a proposed approach that uses the previous decoder's input and hidden states of a model. In this section, we evaluate the differences in the input types and the effectiveness of the models. The iTTS's speech quality is objectively evaluated for the differences in input and method.

Table 2 shows the results of the objective evaluations with methods, input features, models, and synthesis chunks. We made four observations regarding the Japanese iTTS. First, the objective evaluation in F0 and MCD demonstrated that the proposed iTTS system approaches the sentence-based TTS (see (3-1) vs. (3-3) or (3-1) vs. (3-4)). Second, the proposed Japanese iTTS systems with **dec+in+hidden** are more efficient than word-based iTTS systems that we reimplemented ourselves. Third, regarding the different input types in the accent phrase, **Pho+AccFeats** is more efficient than **Pho+AccType**. Finally, our proposed method **dec+in+hidden** is better than **dec+in**, that is, using a large amount of previous information is efficient.

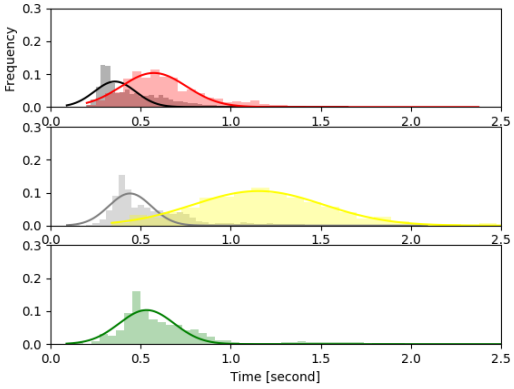



**FIGURE 5.** Relationship between latency and its frequency. The top figure is (1-2): baseline prefix-to-prefix iTTS with one morphoneme+look-*1* (black: average latency of 0.464 seconds) and (3-3): iTTS with one accent phrase as a synthesis chunk (red: average latency of 0.655 seconds). The middle figure is (1-3): baseline prefix-to-prefix iTTS with one morphoneme+look-*2* (gray: average latency of 0.572 seconds) and (3-4): iTTS with two accent phrases as a synthesis chunk (yellow: average latency of 1.20 seconds). The bottom figure is (1-4):prefix-to-prefix iTTS with one morphoneme+look-*3* (green: average latency of 0.667 seconds).

### C. OBJECTIVE EVALUATION OF RELATIONSHIP BETWEEN SPEECH QUALITY AND LATENCY

An accent phrase is longer than a morpheme. Therefore, we must compare the latency of the baseline iTTS and our proposed iTTS before subjectively evaluating our proposed method. We analyzed the latency of the iTTS models with five methods: (1-2), (1-3), (1-4), (3-3), and (3-4).

Fig. 5 shows latencies and their frequencies. The latencies of the proposed iTTS were 0.655 seconds with one accent phrase and 1.20 seconds with two accent phrases. While the latencies of the baseline methods are 0.572 seconds with look-ahead of two morphemes and 0.667 seconds with look-ahead of three morphemes. The latency of the proposed method with one accent phrase is slightly slower but has lower F0 error and MCD.

**TABLE 2. Objective evaluations of proposed methods. Note that full sentence in a column of the synthesis chunk means a sentence unit as the synthesis chunk, 1 morpheme+look-*k* means one morpheme as the synthesis chunk with a look-ahead of *k* length, and 1 accent phrase and 2 accent phrases mean one accent phrase as the synthesis chunk and two accent phrases as the synthesis chunk.**

| Methods | Models | Synthesis chunk | F0 error [cent] | MCD [dB] |
|---|---|---|---|---|
| Input type: **Pho** | | | | |
| (1-1): Baseline TTS | Sentence-based TTS | Fulll sentence | 276.98 | 7.03 |
| (1-2): Baseline prefix-to-prefix iTTS | Word-based iTTS | 1 morpheme+look-*1* | 318.99 | 7.24 |
| (1-3): Baseline prefix-to-prefix iTTS | Word-based iTTS | 1 morpheme+look-*2* | 317.38 | 7.20 |
| (1-4): Baseline prefix-to-prefix iTTS | Word-based iTTS | 1 morpheme+look-*3* | 311.162 | 7.20 |
| (1-5): Baseline prefix-to-prefix iTTS | Word-based iTTS | 1 morpheme+look-*4* | 303.001 | 7.21 |
| (1-6): Baseline prefix-to-prefix iTTS | Word-based iTTS | 1 morpheme+look-*5* | 304.235 | 7.14 |
| (1-7): Baseline prefix-to-prefix iTTS | Word-based iTTS | 1 morpheme+look-*10* | 290.438 | 7.08 |
| Input type: **Pho+AccType** | | | | |
| (2-1): Baseline TTS | Sentence-based TTS | Full sentence | 237.10 | 6.87 |
| (2-2): Japanese iTTS | Accent phrase-based iTTS with dec+in | 1 accent phrase | 309.57 | 7.32 |
| (2-3): Japanese iTTS | Accent phrase-based iTTS with dec+in+hidden | 1 accent phrase | 273.04 | 7.27 |
| (2-4): Japanese iTTS | Accent phrase-based iTTS with dec+in+hidden | 2 accent phrases | 278.34 | 7.12 |
| Input type: **Pho+AccFeats** | | | | |
| (3-1): Topline TTS | Sentence-based TTS | Full sentence | **212.81** | 6.84 |
| (3-2): Japanese iTTS | Accent phrase-based iTTS with dec+in | 1 accent phrase | 290.84 | 7.33 |
| (3-3): Japanese iTTS | Accent phrase-based iTTS with dec+in+hidden | 1 accent phrase | **246.16** | 7.08 |
| (3-4): Japanese iTTS | Accent phrase-based iTTS with dec+in+hidden | 2 accent phrases | **230.03** | 6.94 |

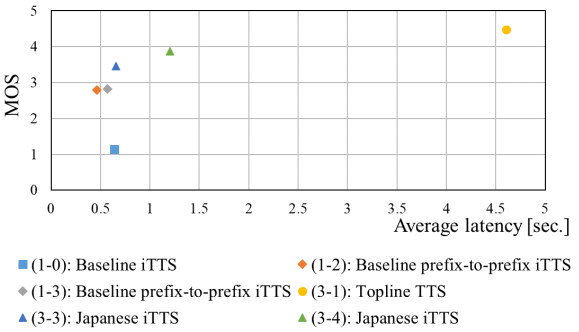



**FIGURE 6. Relationship between MOS and its average latency. Note that MOS score of (1-0) is 1.14, (1-2) is 2.78, (1-3) is 2.83, (3-1) is 4.47, (3-3) is 3.45, and (3-4) is 3.87.**

### D. SUBJECTIVE EVALUATION OF NATURALNESS

Next, we conducted a MOS test as a subjective evaluation of naturalness under six experimental conditions: a baseline iTTS with one morpheme as a synthesis chunk and no look-ahead approach (1-0), a baseline prefix-to-prefix iTTS with one morpheme as a synthesis chunk and a look-ahead-*1* (1-2), a baseline prefix-to-prefix iTTS with one morpheme as a synthesis chunk and a look-ahead-*2* (1-3), a topline TTS (3-1), our proposed iTTS with one accent phrase as a synthesis chunk (3-3), and our proposed iTTS with two accent phrases as a synthesis chunk (3-4). Method (1-0) used one morpheme as a short sentence with the baseline TTS model.

The subjective evaluation results and each average latency are shown in Fig. 6. The baseline iTTS that utilized each morpheme as a short sentence showed a lower quality than the baseline prefix-to-prefix iTTS systems, since that model did not use look-ahead inputs [22]. The best model was the topline TTS with all phonemes and accent features in the sentence; the latency was 4.63 seconds and longer than the others. Although the objective results showed that the proposed iTTS system approached the level of the topline TTS, the subjective results of the proposed iTTS showed room for improvement. The intonation between accent phrases might attract strong attention from the evaluators, or the Japanese iTTS system might need more comprehensive information.

iTTS systems with accent phrases showed better quality than baseline iTTS systems. Furthermore, a statistical significance test was conducted between the baseline prefix-to-prefix iTTS with look-ahead-*2* and each proposed Japanese iTTS, and a significant difference was confirmed ($p<0.001$).

## V. CONCLUSION

This paper proposed a novel Japanese end-to-end neural iTTS architecture using an accent phrase unit. We presented a method to connect the initial input by considering the acoustic time-series well as a method to connect the model's internal state. Moreover, we used two types of input features for the accent phrase unit. We experimentally investigated the latency of various iTTS systems with different modeling and synthesis chunks.

We objectively evaluated the speech quality regarding the differences in input and method. The objective evaluation in F0 and MCD demonstrate that the proposed iTTS system approaches the sentence-based TTS, while the MOS score of the proposed iTTS is still lower. The proposed Japanese iTTS systems using the previous initial input and the previous model's internal state are more efficient than word-based iTTS systems. The speech quality is improved by using many features in the accent phrase unit for inputs. Moreover, using a large amount of previous information is also efficient.

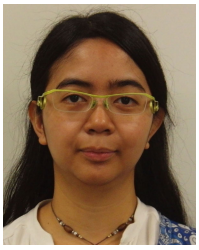
Futhermore, we also subjectively evaluated speech quality. Our results reveal that the proposed method with one accent

phrase had better MOS scores, with a similar latency range between the baseline with look-ahead of two morphemes and the baseline with look-ahead of three morphemes. A method with two accent phrases improved speech quality, although the latency is slightly longer than in a baseline with a two-morpheme look-ahead approach.

## REFERENCES

[1] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR multilingual speech-to-speech translation system," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 2, pp. 365–376, Mar. 2006.
[2] F. Goldman-Eisler, "Segmentation of input in simultaneous translation," *J. Psycholinguistic Res.*, vol. 1, no. 2, pp. 127–140, 1972.
[3] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, Aug. 2017, pp. 4006–4010.
[4] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. ICASSP*, Apr. 2018, pp. 4779–4783.
[5] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, robust and controllable text to speech," in *Proc. NeurIPS*, vol. 32, 2019, pp. 1–10.
[6] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2021, pp. 1–15.
[7] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, "Flow-TTS: A non-autoregressive network for text to speech based on flow," in *Proc. ICASSP*, May 2020, pp. 7209–7213.
[8] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Speech Synth. Workshop*. USA: IEEE Santa Monica, Sep. 2002, pp. 227–230.
[9] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, May 2013, pp. 7962–7966.
[10] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. ICASSP*, Apr. 2015, pp. 4470–4474.
[11] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *Proc. 9th Speech Synth. Workshop (Interspeech)*, 2016, p. 125.
[12] A. Oord et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3918–3926.
[13] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, May 2019, pp. 3617–3621.
[14] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2410–2419.
[15] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, May 2020, pp. 6199–6203.
[16] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33. Red Hook, NY, USA: Curran Associates, 2020, pp. 17022–17033.
[17] T. Baumann and D. Schlangen, "Evaluating prosodic processing for incremental speech synthesis," in *Proc. Interspeech*, Sep. 2012, pp. 438–441.
[18] M. Pouget, T. Hueber, G. Bailly, and T. Baumann, "HMM training strategy for incremental speech synthesis," in *Proc. Interspeech*, Sep. 2015, pp. 1201–1205.
[19] M. Pouget, O. Nahorna, T. Hueber, and G. Bailly, "Adaptive latency for part-of-speech tagging in incremental text-to-speech synthesis," in *Proc. Interspeech*, Sep. 2016, pp. 2846–2850.
[20] T. Yanagita, S. Sakti, and S. Nakamura, "Incremental TTS for Japanese language," in *Proc. Interspeech*, Sep. 2018, pp. 902–906.
[21] T. Yanagita, S. Sakti, and S. Nakamura, "Neural iTTS: Toward synthesizing speech in real-time with end-to-end neural text-to-speech framework," in *Proc. 10th ISCA Workshop Speech Synth. (SSW)*, Sep. 2019, pp. 183–188.
[22] M. Ma, B. Zheng, K. Liu, R. Zheng, H. Liu, K. Peng, K. Church, and L. Huang, "Incremental text-to-speech synthesis with prefix-to-prefix framework," in *Proc. Findings Assoc. Comput. Linguistics EMNLP*, 2020, pp. 3886–3896.
[23] B. Stephenson, L. Besacier, L. Girin, and T. Hueber, "What the future brings: Investigating the impact of lookahead for incremental neural TTS," in *Proc. Interspeech*, Oct. 2020, pp. 215–219.
[24] D. S. R. Mohan, R. Lenain, L. Foglianti, T. H. Teh, M. Staib, A. Torresquintero, and J. Gao, "Incremental text to speech for neural sequence-to-sequence models using reinforcement learning," in *Proc. Interspeech*, Oct. 2020, pp. 3186–3190.
[25] B. Stephenson, T. Hueber, L. Girin, and L. Besacier, "Alternate endings: Improving prosody for incremental neural TTS with predicted future text input," in *Proc. Interspeech*, Aug. 2021, pp. 3865–3869.
[26] T. Saeki, S. Takamichi, and H. Saruwatari, "Incremental text-to-speech synthesis using pseudo lookahead with large pretrained language model," *IEEE Signal Process. Lett.*, vol. 28, pp. 857–861, 2021.
[27] S. Yokomizo, T. Nose, and T. Kobayashi, "Evaluation of prosodic contextual factors for HMM-based speech synthesis," in *Proc. Interspeech*, 2010, pp. 430–433.
[28] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *Proc. ICASSP*, May 2019, pp. 6905–6909.
[29] T. Fujimoto, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Impacts of input linguistic feature representation on Japanese end-to-end speech synthesis," in *Proc. 10th ISCA Workshop Speech Synth.*, Sep. 2019, pp. 166–171.
[30] K. Kurihara, N. Seiyama, and T. Kumano, "Prosodic features control by symbols as input of sequence-to-sequence acoustic modeling for neural TTS," *IEICE Trans. Inf. Syst.*, vol. E104.D, no. 2, pp. 302–311, 2021.
[31] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis," 2017, *arXiv:1711.00354*.
[32] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Forward attention in sequence-to-sequence acoustic modeling for speech synthesis," in *Proc. ICASSP*, Apr. 2018, pp. 4789–4793.
[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, Y. Bengio and Y. LeCun, Eds., 2015, pp. 1–15.
[34] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Commun. Comput. Signal Process.*, vol. 1, May 1993, pp. 125–128.
[35] *Methods for Subjective Determination of Transmission Quality*, document ITU-T P.800, 1996.

**TOMOYA YANAGITA** received the master's degree from the Nara Institute of Science and Technology (NAIST) in 2018. He graduated from a doctoral program at NAIST in March 2022. He is currently a Researcher with the Augmented Human Communication Laboratory, NAIST. His research interests include text-to-speech synthesis, deep learning, and incremental speech processing for simultaneous speech translation systems. He is a member of ASJ and IPSJ.



**SAKRIANI SAKTI** (Member, IEEE) received the B.E. degree (cum laude) in informatics from the Bandung Institute of Technology, Indonesia, in 1999, and the M.Sc. and Ph.D. degrees from the University of Ulm, Germany, in 2008 and 2002, respectively. During her thesis work, she worked with the Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. In 2000, she received the DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm. Between 2003 and 2009, she worked as a Researcher at ATR SLC Labs, Japan, and during 2006 to 2011, she worked as an Expert Researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan. In 2005 to 2008, she continued her studies with Dialog Systems Group, University of Ulm. She has been actively involved in collaboration activities such as Asian Pacific Telecommunity Project 2003 to 2007, A-STAR, and USTAR 2006 to 2011. In 2009 to 2011, she served as a Visiting Professor at the Computer Science Department, University of Indonesia (UI), Indonesia. In 2011 to 2017, she was an Assistant Professor at the Augmented Human Communication Laboratory, NAIST, Japan. She also served as a Visiting Scientific Researcher of INRIA Paris-Rocquencourt, France, in 2015 to 2016, under the JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation. In 2018 to 2021, she was a Research Associate Professor at NAIST and a Research Scientist at RIKEN, Center for Advanced Intelligent Project (AIP), Japan. Currently, she is an Associate Professor at JAIST, an Adjunct Associate Professor at NAIST, a Visiting Research Scientist at RIKEN AIP, and also an Adjunct Professor at the University of Indonesia. She is a member of JNS, SFN, ASJ, ISCA, and IEICE. She is also a Committee Member of IEEE SLTC (2021–2023) and an Associate Editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2020–2023). She is also the chair of ELRA/ISCA Special Interest Group on Under-resourced Languages (SIGUL) and a Board Member of Spoken Language Technologies for Under-Resourced Languages (SLTU). Her research interests include statistical pattern recognition, graphical modeling framework, deep learning, multilingual speech recognition and synthesis, spoken language translation, affective dialog system, and cognitive-communication.

**SATOSHI NAKAMURA** (Fellow, IEEE) received the B.S. degree from the Kyoto Institute of Technology, in 1981, and the Ph.D. degree from Kyoto University, in 1992. He was an Associate Professor with the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), from 1994 to 2000, the Director of ATR Spoken Language Communication Research Laboratories, from 2000 to 2008, and the Vice President of ATR, from 2007 to 2008. He was the Director General of Keihanna Research Laboratories and the Executive Director of the Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan, from 2009 to 2010, and a Project Leader of the Tourism Information Analytics Team, Center for Advanced Intelligence Project (AIP), RIKEN Institute, from 2017 to 2021. He is currently the Director of the Augmented Human Communication Laboratory and a Full Professor with the Graduate School of Information Science, NAIST. He is a leading Researcher of speech-to-speech translation and has been participating in various worldwide speech-to-speech translation research projects, including C-STAR, IWSLT, and A-STAR. His research interests include modeling and systems of speech-to-speech translation and speech recognition. He was elected as a Board Member of the International Speech Communication Association, from 2011 to 2018, an Editorial Board Member of *IEEE Signal Processing Magazine*, from 2012 to 2014, and an IEEE SPS Speech and Language Technical Committee Member, from 2013 to 2015. He is an ISCA Fellow, IPSJ Fellow, and ATR Fellow. He received the Yamashita Research Award, the Kiyasu Award from the Information Processing Society of Japan, the Telecom System Award, the AAMT Nagao Award, the Docomo Mobile Science Award in 2007, the ASJ Award for Distinguished Achievements in Acoustics, the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, the Commendation for Science and Technology by the Minister of Internal Affairs and Communications, and the LREC Antonio Zampolli Prize, in 2012.