Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Excitement-Inducing Commentary Text-to-Speech System for Fighting Game Video Scenes

KOTA IURA[1], (NonMember), YUKI SAITO[1], (Member, IEEE), SHINNOSUKE TAKAMICHI[1,2], (Member, IEEE), GRAHAM NEUBIG[3], (Member, IEEE), KATSUHITO SUDOH[4], (Nonmember), HIROSHI SARUWATARI[1], (Member, IEEE), HIROYA TAKAMURA[5], (Member, IEEE), TATSUYA ISHIGAKI[5], (Member, IEEE)

[1]The University of Tokyo, Hongo, Bunkyo-ku, Tokyo, Japan
[2]Keio University, Hiyoshi, Kohoku-ku, Yokohamama, Kanagawa, Japan
[3]Carnegie Mellon University, Forbes Avenue, Pittsburgh, PA, 15213, U.S.A.
[4]Nara Women's University, Kitauoya Nishimachi, Nara, Japan
[5]National Institute of Advanced Industrial Science and Technology, Aomi, Koto-ku, Tokyo, Japan

Corresponding author: Yuki Saito (yuuki_saito@ipc.i.u-tokyo.ac.jp).

This work was supported by JSPS KAKENHI, Grant Number 11K17945. This paper is based on results obtained from a project, Programs for Bridging the gap between R&D and the IDeal society (society 5.0) and Generating Economic and social value (BRIDGE)/Practical Global Research in the AI × Robotics Services, implemented by the Cabinet Office, Government of Japan.

ABSTRACT In recent years, video games have become a spectator activity, with e-sports and live streaming attracting large audiences. In e-sports, human commentators can enhance viewer excitement by accurately describing match situations and adjusting their speaking style to match exciting scenes. While automation of text generation from video games has been researched, generation of commentary speech that excites the audience has received little attention. To this end, we focus on modeling an e-sports commentator's speaking style. Specifically, we propose a text-to-speech (TTS) system that can adjust the excitement level of synthetic speech on the basis of the gameplay videos. Our system consists of a TTS model and an excitement score predictor. The TTS model is trained to predict commentary speech from text and a binary excitement label. The excitement score predictor rates the excitement level of gameplay videos and feeds the binarized excitement score to the TTS model. We validate the effectiveness of our system using gameplay videos of Super Smash Bros. Ultimate. Subjective evaluations demonstrate that 1) our TTS model effectively controls excitement levels and 2) the TTS system enhanced the audience's engagement and entertainment compared to a baseline TTS system.

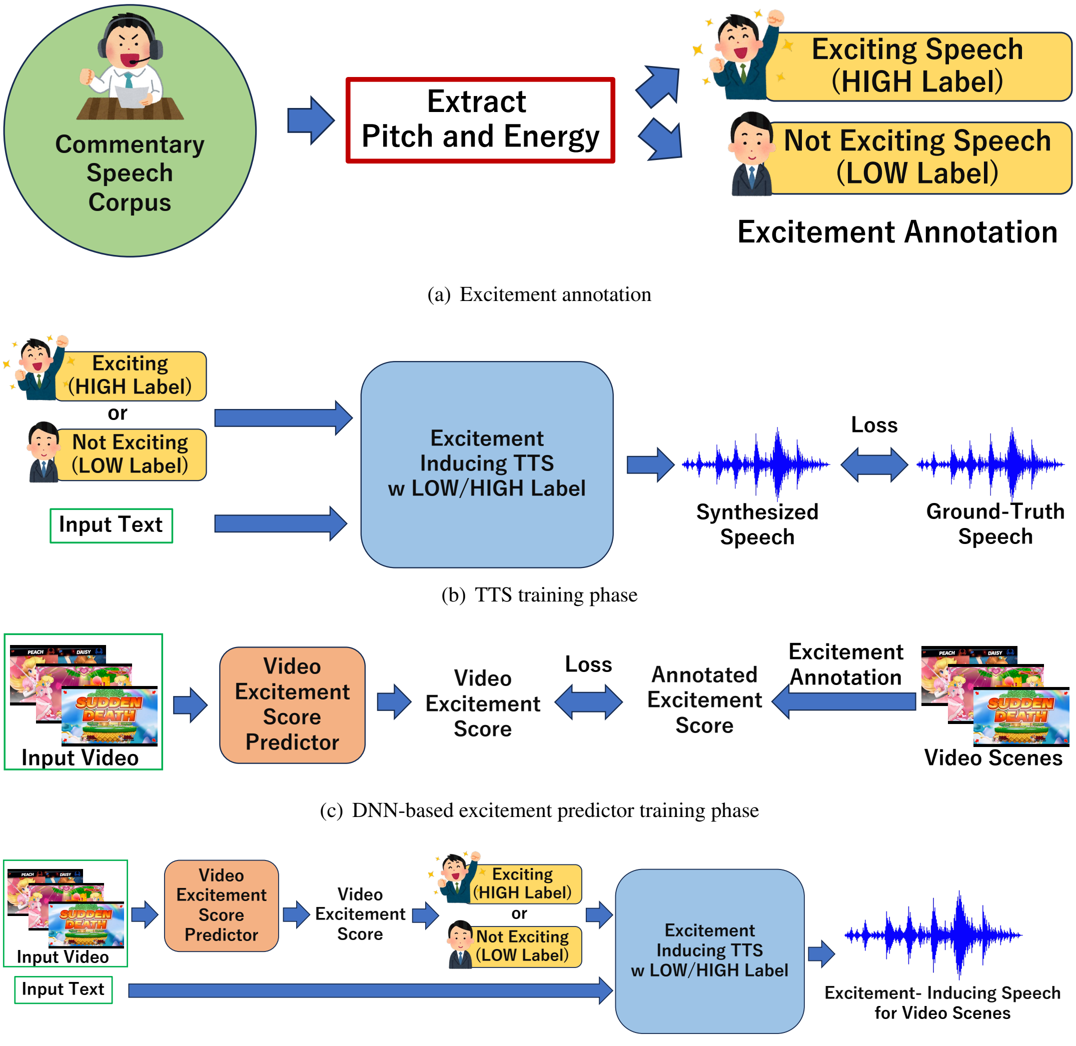
INDEX TERMS TTS, video games, fighting games, e-sports, commentary speech

I. INTRODUCTION

In recent years, video games have evolved beyond just being played—an increasing number of people enjoy watching gameplay through e-sports and live streaming [1, 2]. In this context, human commentators play a crucial role by accurately describing game scenes and adjusting their speaking style to both inform and excite the audience [3, 4]. Thus, commentary speech is essential in making game spectating more enjoyable.

However, delivering effective commentary speech requires in-depth knowledge of the game and players, as well as the ability to instantly describe gameplay scenes and appropriately control their speaking style [4]. Due to these demanding requirements, securing a sufficient number of skilled commentators remains a challenge. To address this issue, automated commentary systems have been actively researched [5, 6, 7]. Although studies on commentary *text* generation have advanced thanks to deep learning technologies, synthesizing expressive commentary *speech*—for example, raising pitch or accelerating speech to convey excitement—has yet to be sufficiently explored.

To this end, we aim to develop a text-to-speech (TTS) system that can change speaking-style on the basis of visual observation from the gameplay scenes. Specifically, we focus on live commentary for fighting games for three main reasons: 1) the availability of a well-structured commentary speech corpus [8], 2) the high-paced scene changes that make the speaking-style control more challenging, and 3) the significant speaking-style variations among human commentators. As shown in Figure 1, our TTS system consists of an excitement-inducing TTS model and a video excitement score predictor. First, to build our system, we annotate ex-



(a) Excitement annotation

(b) TTS training phase

(c) DNN-based excitement predictor training phase

(d) Inference phase

FIGURE 1: Overview of the commentary TTS system with excitement control based on video scenes

citement scores to actual commentary speech from a publicly available corpus and analyze the distribution of the scores on the basis of the pitch and energy of commentary speech to make binary excitement labels (Figure 1(a)). Second, we train the TTS model using the commentary speech corpus with the excitement labels to minimize a loss function between ground-truth and synthesized speech samples (Figure 1(b)). Third, to construct the predictor that estimates the excitement level of an input gameplay scene during inference, we also annotate excitement scores to gameplay videos in the commentary speech corpus. The predictor is trained to minimize a loss function between annotated video excitement scores and predicted ones from the input gameplay videos (Figure 1(c)). Finally, we concatenate the predictor with the TTS model to construct our TTS system that can control the excitement of synthetic speech by using the binarized video excitement score (HIGH or LOW) predicted from the score predictor.

In an experimental evaluation using the SMASH corpus [8], we first verify that the TTS model accurately reproduces excitement levels controlled by given excitement labels. We then assessed whether the full system can generate exciting and engaging commentary speech. Results showed that our system achieved higher excitement ratings than a baseline system unconditioned by the excitement labels.

The rest of our paper is organized as follows. Section II reviews previous work related to our paper. Section III describes the proposed TTS model. Section IV explains the excitement score predictor. Section V presents our experiments and discussions Section VI concludes our paper with future work.

II. RELATED WORK

A. *EXPRESSIVE TTS*

Advanced expressive TTS can control speech emotion [9, 10]. These methods typically require pre-defined labels to control synthetic speech in inference for reproducing desired speaking styles learned during TTS model training.

However, these methods generally require emotion labels to be explicitly input, so they lack the improvisational adaptability to dynamically adjust speaking styles on the basis of surrounding contexts the way humans do. To address this issue, we aim to develop a TTS system able to adjust speaking styles in accordance with video scenes. We consider the perceived excitement of gameplay scenes as an alternative input to control synthetic speech in TTS. In addition, we investigate a way to realize the improvisational control of synthetic speech in accordance with excitement labels predicted from the gameplay scenes.

B. AUTOMATIC COMMENTARY SYSTEM

In sports and e-sports, commentators excite audiences by explaining the game situation, helping the audiences better understand the matches [3, 4]. To emulate such commentators' behaviors through computers and artificial intelligences, many studies have been conducted on live commentary generation. Commentary text generation has been researched in various fields, including sports [11, 12, 13, 14], board games [15, 16], and e-sports [5, 6]. Most of these studies focus on generating commentary text on the basis of structured data or video input. However, these studies primarily emphasize generating appropriate commentary text and determining suitable timing for commentary, without considering speech synthesis for audio modality. Although some previous work [17, 7] addressed generation of commentary speech, their TTS models synthesize reading-style speech, which is insufficient to excite audiences. In this study, we focus on synthesizing commentary speech that enhances audience excitement.

C. AUDIOBOOK SYNTHESIS

As expressive TTS methods (Section II-A) have progressed, they are increasingly been applied for entertainments. One well-known application is audiobook synthesis, where the TTS model incorporates a text encoder (e.g., BERT [18]) to predict the speech style from the input audiobook text. The TTS model synthesizes audiobook speech using the predicted speech style. Our TTS system was inspired by this, and we propose a new TTS model with speaking-style prediction from gameplay video scenes.

D. SMASH CORPUS

To facilitate research on e-sports understanding, Saito et al. constructed the SMASH corpus [8] featuring Super Smash Bros. Ultimate (SSBU)[1]. The SMASH corpus contains Japanese commentary speech recorded over SSBU gameplay videos with transcriptions and timestamps provided for the commentary. In this study, we utilize this corpus to investigate the relationship between perceived excitement, speech features, and video scenes.

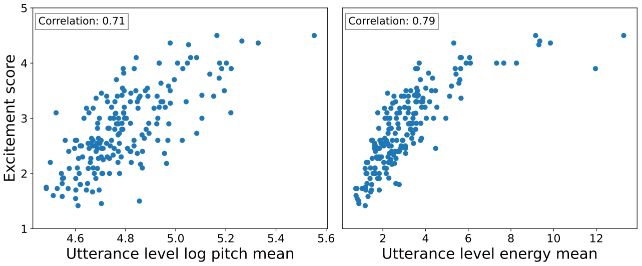


FIGURE 2: Scatter plots of excitement scores against log pitch (left) and energy (right) of commentary speech

III. EXCITEMENT-INDUCING TTS MODEL

This section describes our proposed excitement-inducing TTS model. First, we investigate the relationship between excitement and prosodic features of commentary speech. Then, we explain the labeling method for excitement control. Finally, we provide an overview of the proposed TTS model.

A. COLLECTION OF EXCITEMENT SCORES FOR COMMENTARY SPEECH

We first examined the relationship between excitement and speech features. We conducted a subjective evaluation using the MC1 commentator in SMASH [8] as stimuli to analyze this relationship. We selected 191 utterances (approximately 18 minutes) from six matches. A total of 100 listeners answered excitement scores of 20 randomly selected utterances on a five-point scale (1: not exciting – 5: very exciting) via the crowdsourcing platform Lancers[2].

To examine prosodic features that affect excitement perceived by listeners, we analyzed the relationship between speech excitement scores, pitch, and energy. Pitch was extracted using WORLD [19] (D4C edition [20]) with a sampling rate of 22,050 Hz. Energy was computed from the amplitude spectrum obtained by short-time Fourier transform, setting its frame size, hop size, and window length to 1,024, 256, and 1,024, respectively. The norm of each frame's amplitude spectrum was used as the energy measure.

As shown in Figure 2, pitch and energy strongly correlate with speech excitement scores, with their Pearson correlation coefficients of 0.71 and 0.79, respectively. This highlights their importance in synthesizing speech with appropriate excitement levels.

B. EXCITEMENT LABELING BASED ON PROSODIC FEATURES

Our analysis in Section III-A revealed that pitch and energy are essential factors that affect the excitement scores of commentary speech. Therefore, we assign a speech excitement label for the MC1 commentator's each utterance, on the basis of threshold values for the pitch and energy. Specifically, we first normalized pitch and energy values to have zero mean and unit variance. Then, we set the threshold values for pitch

[1] https://www.smashbros.com/en_US/index.html

[2] https://www.lancers.jp/

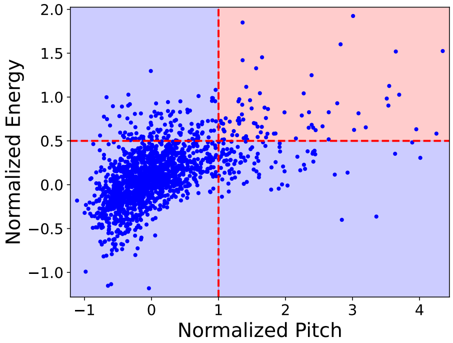


FIGURE 3: Scatter plot of mean normalized pitch and energy for each utterance. The red lines indicate the pitch and energy thresholds used for "LOW" or "HIGH" label assignment.

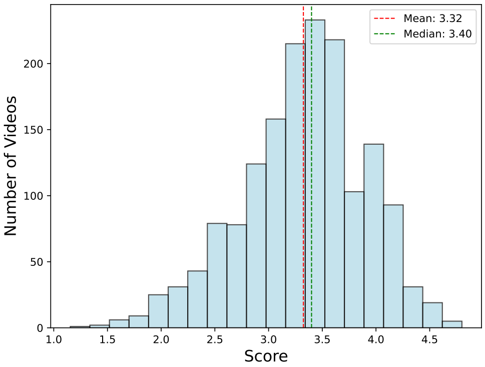


FIGURE 4: Histogram of average video excitement scores on the SMASH corpus

and energy to 1.0 and 0.5, respectively. Finally, we classified utterances that surpassed both the thresholds as "HIGH" excitement and others as "LOW" excitement. Figure 3 shows the scatter plot of normalized pitch and energy with their threshold values. This labeling resulted in 73 out of 1,611 utterances from MC1 in the SMASH corpus being labeled as "HIGH" and the remaining 1,538 as "LOW."

C. TTS MODEL FOR INDUCING EXCITEMENT USING LABELS

We use VITS [21] and JETS [22] as backbone TTS models. Both are end-to-end TTS models that enable fast and stable training even on relatively small datasets. VITS is widely used as a baseline due to its high speech naturalness. JETS includes a pitch/energy predictor similar to the FastSpeech 2-based TTS model [23], which is expected to offer better controllability in these prosodic factors.

We implement the VITS-based and JETS-based TTS models following the ESPnet framework [24] with conditioning by an excitement label converted into trainable embedding (i.e., excitement embeddings). In VITS, following the style embedding method from PromptStyle [25], we input the excitement embeddings to three modules: the posterior encoder, flow, and stochastic duration predictor. In JETS, following the speaker embedding approach in ESPnet[3], we add the excitement embeddings to the text embeddings predicted by the text encoder.

IV. EXCITEMENT LABEL PREDICTION FROM VIDEO

When controlling speech excitement on the basis of video, a natural approach is to predict excitement labels or *speech* excitement scores from the video directly. However, one can introduce *video* excitement scores into our TTS system because the process of the threshold adjustment for "LOW"/"HIGH" excitement prediction using speech features should be highly commentator-dependent. Therefore, we investigate alternative ways to determine the excitement level from video scenes and use this information to control the TTS input labels.

A. EXCITEMENT ANNOTATION BASED ON VIDEO-ONLY PERCEPTION

First, to assess the subjective excitement of gameplay scenes, we annotate excitement scores on the videos included in the SMASH corpus. To identify exciting video scenes, we conducted an annotation experiment where participants watched only the gameplay videos from the SMASH corpus and rated each video scene's excitement score on a five-point scale. We first segmented the videos on the basis of the speech utterance timestamps in the SMASH corpus, which resulted in 1,611 video segments. To make participants rate the excitement scores solely on the visual content, we removed game sounds and commentary voices from the videos[4]. Annotation was then conducted via Lancers with 780 participants, each rating 21 video segments. Each segment was scored by at least 10 participants using a five-point scale (1: not exciting – 5: very exciting).

Then, we averaged the five-point scores annotated to each video. Figure 4 shows the histogram of the average video excitement scores. The overall mean score was 3.32, and the median was 3.40, indicating that many videos received relatively high excitement scores. This trend likely results from the visual effects in SSBU, which naturally contribute to excitement. Video scenes with scores ≥ 4.5 were match-related, including Final Smash (a powerful special move), knock outs (KOs), and match conclusions, which had a major impact on the game and featured significant visual transitions. In contrast, video scenes with scores ≤ 1.8 consisted of match start or result screens, where there was little visual change related to the match scenes. These findings suggest that, when assessing excitement solely from video, match-related scenes and screen movement intensity play a key role in perceived excitement.

In the following sections, we propose two methods for predicting the excitement score from video. One method determines video scene excitement using hard-crafted video

[3] https://github.com/espnet/espnet

[4] Although the game sound is another important factor that affects the excitement of gameplay scenes, we decided to mute it in this annotation experiment because the SMASH corpus contains the gameplay videos mixed with the commentary voices.

features. The other method utilizes a deep neural network (DNN) model that takes video as input and predicts its excitement score.

B. EXCITEMENT PREDICTION USING HAND-CRAFTED VIDEO FEATURES

From the analysis results of excitement scores derived from video, we observed that video scenes with higher excitement ratings generally involved significant visual changes and intense action. On the basis of this observation, we consider two metrics that can be computed from individual video frames to represent screen changes. By leveraging these metrics, we aim to predict whether a video scene is perceived as exciting.

Let $I_t(m,n)$ represent the pixel intensity at position (m,n) in the grayscale image of the tth frame, where:

- $m = 1, 2, \ldots, M$: Row index of the image (vertical dimension)
- $n = 1, 2, \ldots, N$: Column index of the image (horizontal dimension)

We use this notation as the basis for representing video-based excitement prediction using hand-crafted features.

Frame Difference (FD): Video scenes with significant screen changes are expected to exhibit large pixel variations between consecutive frames. Therefore, we quantify the intensity of screen motion by computing the pixel difference between consecutive frames across the entire image and averaging the values. The metric $\mathrm{FD}(I_t)$, which represents the amount of change in the tth frame, is defined as:

$$\mathrm{FD}(I_t) = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} |I_t(m,n) - I_{t-1}(m,n)|.$$

By utilizing $\mathrm{FD}(I_t)$, we can numerically represent visually intense scenes.

Laplacian Variance (LV): Video scenes perceived as exciting often contain distinctive visual effects, such as explosion blasts from KOs or special attacks from Final Smashes in the SSBU case. To detect such exciting scenes, we utilize LV [27], a measure of image sharpness that is commonly used to assess blurriness. The LV is computed using the following procedure. The Laplacian operation used in LV is expressed as:

$$L_t(m,n) = (I_t * K)(m,n),$$

where $L_t(m,n)$ represents the result of the Laplacian operation, and $I_t * K$ denotes the convolution of the image I_t with the Laplacian kernel K defined as:

$$K = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}.$$

The variance of $L_t(m,n)$ over the entire image is then computed as:

$$\mathrm{LV}(I_t) = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \left(|L_t(m,n)| - \bar{L}_t\right)^2,$$

where $\bar{L}_t$ represents the mean absolute Laplacian value:

$$\bar{L}_t = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} |L_t(m,n)|.$$

Images with sharp focus contain more edges, resulting in a higher $\mathrm{LV}(I_t)$. Conversely, in blurred images where the focus is not sharp, $\mathrm{LV}(I_t)$ becomes smaller. In game images, video scenes with numerous visual effects are expected to contain more edges, enabling exciting scenes to be detected on the basis of this metric.

C. DNN-BASED EXCITEMENT PREDICTION

While video metrics computed from image frames offer fast and efficient calculations, they fail to capture the complex visual features of fighting game videos. To overcome this limitation, we explore a DNN-based approach using Vision Transformer (ViT) [26]. By training ViT to predict excitement scores from video segments, we aim to enhance scene excitement prediction.

The overview of the method for predicting excitement using ViT is shown in Figure 5. First, optical flow is computed for each frame of the input video using the Farneback method [28]. Optical flow represents the motion of each pixel between consecutive frames. Second, both the original image and the optical flow image are fed into the ViT model to obtain their respective latent representations. These representations are subsequently concatenated. Third, the frame-level representations are input into a Transformer encoder [29], enabling the model to learn temporal relationships between frames. Finally, the Transformer encoder output for each frame is averaged to construct the video-level latent representation, which is then passed through a linear layer to predict the excitement score. By binarizing the predicted excitement score, we can obtain the HIGH/LOW excitement label that can control the excitement level of synthetic speech from TTS.

V. EXPERIMENTAL EVALUATION

To verify whether the proposed TTS system can generate exciting and engaging commentary speech, we conducted objective and subjective evaluations. First, we assess whether the proposed TTS model can control the excitement level of the generated commentary speech. Second, we quantify the performance of our excitement prediction methods from video scenes. Finally, we examine whether the full proposed system, i.e., TTS with video-based excitement control, enhances viewer excitement and enjoyment.

A. EXPERIMENTAL CONDITIONS

We describe conditions common to all experiments.

Dataset: We used the SMASH corpus [8] for the experimental evaluation. This corpus contains commentary speech from two commentators, MC1 and MC2. However, since this study assumes a single-speaker TTS system, we used the data from MC1, who has a total recording time of approximately

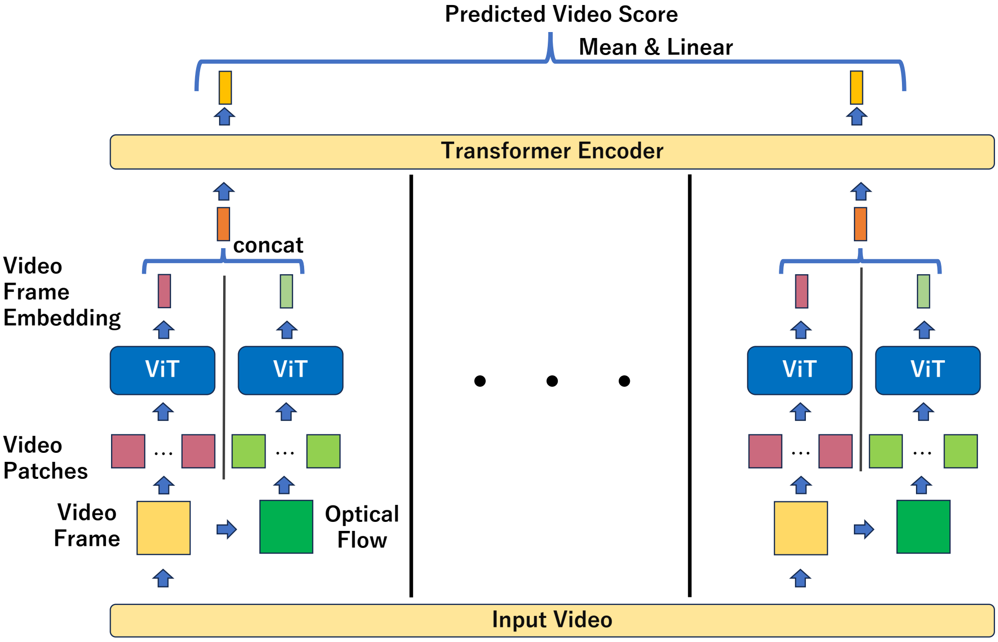


FIGURE 5: Overview of the excitement prediction based on Vision Transformer (ViT) [26].

2.5 hours. We divided this dataset, consisting of 52 matches and 1,611 utterances in total, into three subsets: training (40 matches with 1,234 utterances), validation (6 matches with 186 utterances), and test (6 matches with 191 utterances).

TTS model: To evaluate the effectiveness of the proposed TTS model conditioned on LOW/HIGH labels, we prepared baseline TTS models using VITS [21] and JETS [22] trained without these labels. We used opensourced implementations for VITS and JETS in the ESPnet framework. For VITS, we fine-tuned a pre-trained model[5] trained on the JSUT corpus [30]using the SMASH corpus. The batch size was set to 20, and training was performed for 100 epochs. The top 10 models with the lowest validation loss were selected, and their parameters were averaged to build the final model. For JETS, since pre-trained models with the JSUT corpus were unavailable in ESPnet, we first trained JETS on JSUT with a batch size of 20 for 400 epochs. The top 5 models with the lowest validation loss were selected and averaged. We then fine-tuned the model on the SMASH corpus, training for 300 epochs with a batch size of 20, again averaging the top 5 models for the final model. For all models, AdamW [31] was used as the optimizer, with a learning rate of 10^{-4}, $\beta_1 = 0.8$, and $\beta_2 = 0.99$. Other hyperparameters for VITS and JETS followed the default values used in the ESPnet implementation.

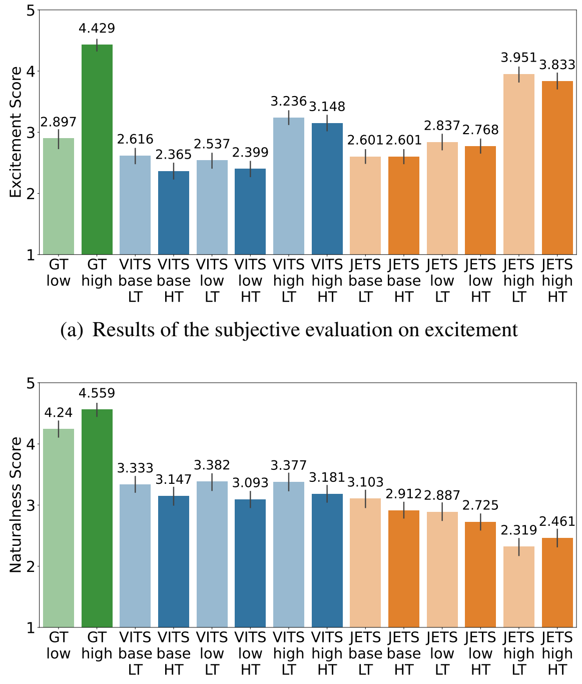
Excitement prediction methods: We implemented two approaches for excitement prediction: one on the basis of video features (VF) and another using a ViT-based model. For VF, we computed FD and LV for each video frame, then averaged these values over the entire video. If either averaged value exceeded a predefined threshold of 3.5 (near the median value), the video was classified as high excitement. To maintain consistency across matches, FD and LV were normalized per match by setting the mean to 0 and variance to 1. For ViT-based prediction, we used `ViT_B_16` from the torchvision library (version 0.17.2), initialized with pre-trained `IMAGENET1K_V1` weights. The model was configured with a patch size of 16, a hidden size of 768, 12 encoder layers, and 12 attention heads. Optical flow was computed using OpenCV's [32] `calcOpticalFlowFarneback` function with default parameters. The two-layer Transformer encoder with a hidden dimension of 768 and 8 attention heads captured temporal relationships between frame embeddings. AdamW was used as the optimizer with a learning rate of 10^{-6}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Training was conducted for 100 epochs. The input image size was set to 224.

B. SUBJECTIVE EVALUATION OF TTS MODELS

To assess whether the proposed TTS model can synthesize speech with acceptable naturalness and controllable excitement, we conducted a subjective evaluation using Lancers with 200 participants. Each participant listened to synthesized speech samples generated from the same text in randomized order and rated them on two aspects with five-point scales: excitement level (1: not exciting – 5: very exciting) and speech naturalness (1: very unnatural – 5: very natural). To compare methods, we performed multiple comparisons using the Steel–Dwass test with a significance level of 0.05.

Subjective Evaluation of Excitement: The results of the subjective evaluation on excitement are shown in Figure 6(a). The ground-truth (GT) speech consists exclusively of either "LOW"-labeled or "HIGH"-labeled speech. Therefore, for

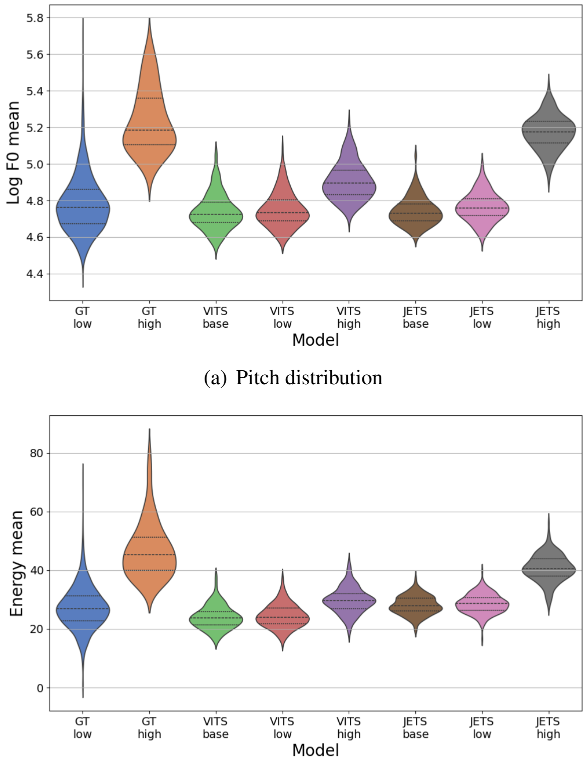
[5] https://zenodo.org/records/5521360



(a) Results of the subjective evaluation on excitement

(b) Results of the subjective evaluation on naturalness

FIGURE 6: Subjective evaluation results for the proposed excitement-controllable TTS model. "LT" and "HT" indicate whether the text of the speech corresponds to LOW and HIGH excitement, respectively. For example, "VITS low HT" means speech samples synthesized with VITS, where the input excitement label is "LOW" and the text is derived from speech annotated with the "HIGH" excitement label. "base" denotes that the TTS models were trained without the excitement label conditioning.



(a) Pitch distribution

(b) Energy distribution

FIGURE 7: Violin plots of pitch and energy

evaluation, we categorized the synthesized speech on the basis of whether the input text originated from "LOW"-labeled or "HIGH"-labeled GT speech. We denote cases where the input text originates from "LOW"-labeled speech as "LT" and those from "HIGH"-labeled speech as "HT." For both JETS and VITS models, the results indicate that the excitement of synthesized commentary voices can be effectively controlled by using the excitement labels. In particular, JETS demonstrates more pronounced control over the excitement levels. Additionally, the baseline models for both VITS and JETS exhibit results similar to the "LOW"-label condition, suggesting that the proposed label-based training enables the TTS model to synthesize speech with higher excitement.

Subjective Evaluation of Naturalness: The results of the naturalness evaluation are shown in Figure 6(b). Both the VITS and JETS models exhibited significantly lower naturalness than GT speech. This can be attributed to the limited amount of training data, which is approximately two hours. Additionally, the SMASH corpus, being a collection of SSBU commentary, contains many proper nouns, including the names of fighters and actions, leading to unnatural accents in the synthesized speech and lower perceived naturalness. Regarding the proposed TTS model, VITS showed naturalness similar to the baseline for both of the "LOW"-label and "HIGH"-label conditions. In contrast, JETS, particularly for the "HIGH"-label condition, exhibited lower naturalness than the baseline. One reason might be that JETS with the "HIGH"-label condition resulted in excessively increased pitch and energy due to the highly imbalanced training data, leading to lower naturalness ratings.

Visualizing the Pitch and Energy Distributions: To verify the excitement controllability of the proposed TTS models by labels, we visualized the pitch and energy distributions for both natural and synthesized speech as violin plots. The results are shown in Figure 7. For both VITS and JETS, the "HIGH"-label condition resulted in higher pitch and energy values, with JETS exhibiting a more pronounced difference than VITS. Additionally, the distributions for JETS were closer to those of natural speech. These findings align with the subjective evaluation results for excitement (Figure 6(a)), demonstrating that the proposed TTS model with label-based control effectively excites the listeners.

C. EVALUATION OF EXCITEMENT PREDICTION METHODS

We evaluated how well video-predicted excitement labels align with speech-derived excitement labels. Considering the imbalance in "HIGH" labels, we used F1-score as an evaluation metric. The F1-scores were 0.17 for the VF-based method and 0.09 for the ViT-based method, indicating low alignment between video-predicted and speech-derived excitement labels. When we change the threshold value for the excitement prediction from 3.5 to 3.8, the score increased to 0.11. One possible reason for this is the severe class imbalance, as only 10 out of 191 test utterances were labeled as "HIGH," making prediction difficult. These results suggest that relying solely on speech-derived "LOW"/"HIGH" labels may be insufficient, and video-based excitement levels could be more effective for control. In the next section, we evaluate whether these video-based excitement prediction methods can excite listeners.

D. EVALUATION OF THE ENTIRE COMMENTARY SYSTEM

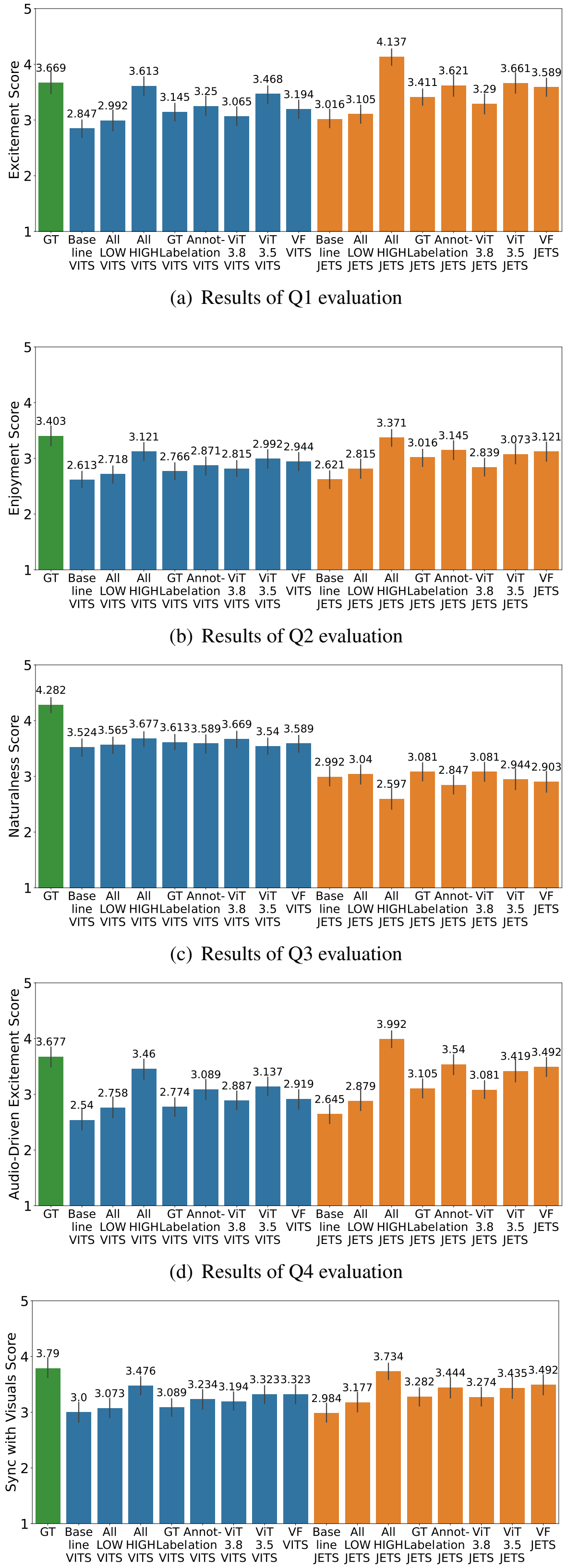
Finally, we evaluated whether the proposed TTS system, which controls excitement on the basis of video scenes, can enhance perceived excitement and entertainment.

Methods for Excitement Control: This evaluation included nine conditions: (1: GT) natural speech by MC1, (2: Baseline) models trained without excitement labels, (3: All HIGH) and (4: All LOW) where all commentary speech was synthesized with "HIGH" or "LOW" labels, (5: GT Label) and (6: Annotation) where the excitement was controlled on the basis of "LOW"/"HIGH" labels from GT speech or from GT gameplay videos, (7: ViT 3.5) and (8: ViT 3.8) DNN-based excitement prediction with different thresholds, and (9: VF) prediction using video-based handcrafted metrics.

Video Scene Selection: A total of 21 scenes were selected, ensuring each contained at least one "HIGH" label on the basis of either "GT Label" or "Annotation" derived from GT data. Video segments began 1.5 seconds before speech onset and ended 1.5 seconds after the natural speech end time, ensuring all synthesized speech fit within the scene. Start times were aligned with the SMASH corpus, and if a synthesized utterance exceeded the available time before the next one, the system forcibly started the next utterance on schedule. Video and synthesized audio were mixed using FFmpeg, with game sound muted so that only commentary was audible. The final video durations ranged from 10 to 33 seconds.

Evaluation Criteria: The evaluation was conducted using the crowdsourcing platform Lancers. A total of 120 participants took part in the evaluation. Each participant was randomly assigned to one scene and watched all 17 methods in a randomized order, rating each one individually. The evaluation included five criteria:

- Q1: Considering the commentary speech, how exciting was the video?
- Q2: Considering the commentary speech, how entertaining was the video?
- Q3: How natural did the commentary speech sound (i.e., how human-like and naturally spoken was it)?
- Q4: Did the commentary speech enhance the excitement of the video?
- Q5: Did the excited speaking style in the commentary speech match the exciting moments in the video?



(a) Results of Q1 evaluation

(b) Results of Q2 evaluation

(c) Results of Q3 evaluation

(d) Results of Q4 evaluation

(e) Results of Q5 evaluation

FIGURE 8: Evaluation results for Q1–Q5. Error bars represent 95% confidence intervals.

Each criterion was rated on a five-point scale for each commentary video. To compare the different methods, multiple comparisons were performed using the Steel–Dwass test with a significance level of 0.05.

Evaluation Results (Q1): The results of the Q1 evaluation are shown in Figure 8(a). Among the proposed methods, VITS-based TTS with "All HIGH" and "ViT 3.5" and JETS-based TTS with "All HIGH," ViT 3.5," and "VF" achieved significantly better scores than the baseline. These results indicate that the proposed TTS systems can synthesize more exciting commentary speech than conventional TTS systems. Furthermore, JETS with "All HIGH" scored significantly better than "GT," demonstrating that the proposed method effectively generates exciting commentary speech.

Evaluation Results (Q2): The results of the Q2 evaluation are shown in Figure 8(b). Among the proposed methods, VITS-based TTS with "All HIGH" and JETS-based TTS with "All HIGH," "ViT 3.5," and "VF" achieved significantly better scores than the baseline systems. These results indicate that the proposed TTS system enhanced viewer enjoyment more effectively than conventional TTS systems.

Evaluation Results (Q3): The results of the Q3 evaluation are shown in Figure 8(c). The synthesized speech, especially from the JETS-based systems, scored lower in naturalness than "GT." This result is consistent with the findings from the speech-only evaluation (Figure 6(b)).

Evaluation Results (Q4): The results of the Q4 evaluation are shown in Figure 8(d). JETS with "All HIGH" achieved significantly better scores, consistent with the results of the Q1 evaluation. Additionally, comparisons with the baseline showed that VITS-based TTS with "All HIGH" and "ViT 3.5" and JETS-based TTS with "All HIGH," "ViT 3.5," and "VF" achieved significantly better scores.

Evaluation Results (Q5): The results of Q5 evaluation are shown in Figure 8(e). VITS-based TTS with "All HIGH" and JETS-based TTS with "All HIGH," "ViT 3.5," and "VF" achieved significantly better scores than the baseline. Among the proposed methods, JETS with "All HIGH" received the highest rating.

Analysis and Discussion of Evaluation Results: Among the proposed systems, "All HIGH" tended to achieve the best results, with ViT-based prediction at a lower threshold performing better. This suggests that a higher number of "HIGH" labels leads to better evaluations. The systems are ranked by "HIGH" label count as follows: All HIGH $>$ ViT 3.5 $>$ Annotation $>$ VF $>$ GT Label $>$ ViT 3.8 $>$ All LOW. This ranking aligns with the Q1 (excitement) and Q2 (enjoyment) scores, indicating that using more "HIGH" labels for conditioning the TTS model resulted in greater perceived excitement and enjoyment. For SSBU commentary—a task where the goal is to enhance audience engagement—consistently using a "HIGH"-excitement speaking style enhances both excitement and entertainment, while "LOW"-excitement styles reduce engagement and should be minimized.

A possible reason for this trend is that MC1 in the SMASH corpus tends to use low-pitch, low-volume voices in his commentary. Effective fighting game commentary requires conveying content with a scene-appropriate speaking style to maintain engagement, likely avoiding unexciting speech. The large proportion of "LOW"-excitement speech in SMASH corpus suggests that listeners perceived parts of the commentary to lack excitement. The proposed TTS system, particularly conditioned on "All HIGH," effectively extracts and learns from the more engaging speech segments in the corpus. By synthesizing commentary voices using "HIGH" labels only, the system could generate more engaging fighting game commentary, improving the audience's overall excitement and entertainment.

VI. CONCLUSION

In this study, we proposed a text-to-speech (TTS) system for fighting game commentary, featuring Super Smash Bros. Ultimate as the subject. In subjective evaluations, the proposed TTS model could effectively control excitement levels and the synthesized commentary enhanced viewers' excitement and enjoyment. For future work, we will investigate a way to improve the naturalness of synthesized commentary, such as adapting a TTS model trained from a larger speech corpus to the commentary TTS [33] and introducing deep learning-based (vision) language models [34] to extract more game-specific knowledge from texts and gameplay videos.

REFERENCES

[1] T. Smith, M. Obrist, and P. Wright, "Live-streaming changes the (video) game," in *Proc. EuroITV*, 2013, p. 131–138.

[2] S. Block and F. Haack, "eSports: a new industry," in *SHS Web of Conferences*, vol. 92, 2021, p. 04002.

[3] J. Bryant, D. Brown, P. W. Comisky, and D. Zillmann, "Sports and spectators: Commentary and appreciation." *Journal of Communication*, vol. 32, no. 1, pp. 109–19, 1982.

[4] L. Li, J. Uttarapong, G. Freeman, and D. Y. Wohn, "Spontaneous, yet studious: Esports commentators' live performance and self-presentation practices," in *Proc. ACM HCI*, vol. 4, no. CSCW2, 2020, pp. 1–25.

[5] T. Tanaka and E. Simo-Serra, "LoL-V2T: Large-scale esports video description dataset," in *Proc. CVPR Workshop*, 2021, pp. 4557–4566.

[6] Z. Wang and N. Yoshinaga, "Commentary generation from data records of multiplayer strategy esports game," in *Proc. NAACL*, 2024, pp. 263–271.

[7] T. Ishigaki, G. Topić, Y. Hamazono, I. Kobayashi, Y. Miyao, and H. Takamura, "Audio commentary system for real-time racing game play," in *Proc. INLG*, 2023, pp. 9–10.

[8] Y. Saito, S. Takamichi, and H. Saruwatari, "SMASH corpus: A spontaneous speech corpus recording third-person audio commentaries on gameplay," in *Proc. LREC*, 2020, pp. 6571–6577.

[9] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Batten-

berg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style Tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, 2018, pp. 5180–5189.

[10] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *Proc. ICASSP*, 2020, pp. 7254–7258.

[11] B. J. Kim and Y. S. Choi, "Automatic baseball commentary generation using deep learning," in *Proc. ACM SAC*, 2020, pp. 1056–1065.

[12] J. Rao, H. Wu, C. Liu, Y. Wang, and W. Xie, "MatchTime: Towards automatic soccer game commentary generation," in *Proc. EMNLP*, 2024.

[13] Y. Taniguchi, Y. Feng, H. Takamura, and M. Okumura, "Generating live soccer-match commentary from play data," in *Proc. AAAI*, vol. 33, no. 01, 2019, pp. 7096–7103.

[14] R. Puduppully and M. Lapata, "Data-to-text generation with macro planning," *Trans. of ACL*, 2021.

[15] A. Sadikov, M. Možina, M. Guid, J. Krivec, and I. Bratko, "Automated chess tutor," in *Computers and Games*, 2007, pp. 13–25.

[16] H. Kameko, S. Mori, and Y. Tsuruoka, "Learning a game commentary generator with grounded move expressions," in *Proc. CIG*, 2015.

[17] T. Kumano, T. Takagi, M. Ichiki, K. Kurihara, H. Kaneko, T. Komori, T. Shimizu, N. Seiyama, A. Imai, and H. Sumiyoshi, "Generation of automated sports commentary from live sports data." in *Proc. BMSB*, 2019, pp. 1–4.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, Jun 2019, pp. 4171–4186.

[19] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[20] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[21] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, 2021, pp. 5530–5540.

[22] D. Lim, S. Jung, and E. Kim, "JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech," in *Proc. Interspeech*, 2022, pp. 21–25.

[23] 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. International Conference on Learning Representations*, 2021.

[24] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[25] G. Liu, Y. Zhang, Y. Lei, Y. Chen, R. Wang, Z. Li, and L. Xie, "PromptStyle: Controllable style transfer for text-to-speech with natural language descriptions," in *Proc. Interspeech*, 2023, pp. 792–796.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[27] J. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia, "Diatom autofocusing in brightfield microscopy: A comparative study," in *Proc. ICPR*, vol. 3, 2000, pp. 314–317.

[28] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. SCIA*, 2003, pp. 363–370.

[29] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[30] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, 2020.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[32] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[33] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. NeurIPS*, Montreal, Canada, Dec. 2018, pp. 4480–4490.

[34] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5625–5644, Aug. 2024.

KOTA IURA received his M.S degree from The University of Tokyo, Japan, in 2025. His research interests include speech synthesis, speech processing, and deep learning.

YUKI SAITO received the Ph.D. degree from The University of Tokyo, Japan, in 2021. He has been a lecturer at The University of Tokyo, Japan, since 2024. His research interests include speech synthesis, voice conversion, and machine learning. He has received more than ten paper awards including the 2020 IEEE SPS Young Author Best Paper Award. He is a member of the Acoustical Society of Japan, IEEE SPS, and Institute of Electronics, Information and Communication Engineers.

SHINNOSUKE TAKAMICHI received the Ph.D. degree from Nara Institute of Science and Technology, Japan, in 2016. He is currently an associate professor of Keio University, Japan. He has received more than 20 paper/achievement awards including the IEEE Signal Processing Society Young Author Best Paper Award and the MEXT Young Scientists' Prize.

GRAHAM NEUBIG received his Ph.D. degree from Kyoto University, Japan, in 2012. He is currently an associate professor at the Language Technology Institute, Carnegie Mellon University, U.S.A. and a chief scientist of All Hands AI. His research interests include multilingual language processing, natural language interfaces to computers, and machine learning for natural language processing.

KATSUHITO SUDOH received his B.S., M.S., and Ph.D. degrees from Kyoto University, in 2000, 2002, and 2015, respectively. After working at NTT Communication Science Laboratories 2002-2017 and Nara Institute of Science and Technology 2017-2024, he has been a professor at Nara Women's University, Japan, since 2024. His research interests include machine translation and natural language processing.

HIROSHI SARUWATARI received his B.E., M.E., and Ph.D. degrees from Nagoya University, Japan, in 1991, 1993, and 2000, respectively. In 1993, he joined SECOM IS Laboratory, Japan, and in 2000, Nara Institute of Science and Technology, Japan. Since 2014, he has been a professor with The University of Tokyo, Japan. His research interests include statistical audio signal processing, blind source separation, and speech enhancement. He put his research into the world's first commercially available independent-component-analysis-based BSS microphone in 2007. He was the recipient of several paper awards from IEICE in 2001 and 2006, from TAF in 2004, 2009, 2012, and 2018, from IEEE-IROS2005 in 2006, and from APSIPA in 2013 and 2018, and also the DOCOMO Mobile Science Award in 2011, Ichimura Award in 2013, Commendation for Science and Technology by the Minister of Education in 2015, Achievement Award from IEICE in 2017, and Hoko-Award in 2018. He has been professionally involved in various volunteer works for IEEE, EURASIP, IEICE, and ASJ.

HIROYA TAKAMURA received his Bachelor's and Master's degrees from the University of Tokyo, Japan, in 1997 and 2000, respectively. He received his Ph.D. from Nara Institute of Science and Technology, Japan, in 2003. He worked as a professor at the Tokyo Institute of Technology, Japan, and is currently a research team leader at AI Research Center of National Institute of Advanced Industrial Science and Technology (AIST), Japan. His current research interests include computational linguistics, especially natural language generation.

TATSUYA ISHIGAKI received his M.E. and Ph.D. degrees from the Tokyo Institute of Technology, Japan, in 2014 and 2019, respectively. He is currently a researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. His research interests lie in natural language processing, with a particular focus on text summarization and natural language generation from multimodal data. He is a journal editor of the Association for Natural Language Processing, Japan. He served as a local chair of INLG 2024 and received the Best Demo Paper Award at INLG 2023.