IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 January 2024, accepted 11 February 2024, date of publication 13 February 2024, date of current version 23 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3365948

RESEARCH ARTICLE

Breaking News System of At-Bat Results From Sports Commentary via Speech Recognition

RIKU IKEDA[1], (Graduate Student Member, IEEE), KAZUMA SAKAMOTO[2], AND YOSHIHIRO UEDA[2]

[1]Graduate School of Sustainable Systems Science, Komatsu University, Komatsu, Ishikawa 923-8511, Japan
[2]Faculty of Production Systems Engineering and Sciences, Komatsu University, Komatsu, Ishikawa 923-8511, Japan

Corresponding author: Riku Ikeda (22211002@komatsu-u.ac.jp)

ABSTRACT The proliferation of over-the-top (OTT) systems has led to a tremendous amount of video content over the Internet that users can watch them regardless of time or location restrictions. For sports games, highlight videos allow users to check the results of games without watching or listening in real-time. Despite these advances, the number of OTT companies broadcasting sports games in real-time is rising. This demonstrates that sports games are valuable video content to view in real-time. However, sports games generally last long, making continuous viewing difficult. To address this issue, X's posts have been employed to support sports viewing to users. However, X's limits on reading and posting and its introduction of paid features, an alternative to X need to be found. In this paper, we propose a novel approach as an alternative to X, using sports commentary to keep track of the state of a game in real-time. In order to classify players' at-bat results from sports commentary, we combine pre-trained speech recognition and language models. In experiments, we used various pre-trained speech recognition models to create a dataset for fine-tuning, fine-tuned several pre-trained language models using this dataset, and compared the models to identify the best combination of pre-trained speech recognition and language models for classifying at-bat results. Moreover, by using both the speech recognition models and the fine-tuned language models, we compared in terms of real-time performance and classification accuracy across multiple games to verify their effectiveness.

INDEX TERMS Real-time systems, at-bat results classification, commentary, sports viewing support, combination of pre-trained speech recognition and language models.

I. INTRODUCTION

The rise of smartphones and increased communication speed have led to the advent of numerous over-the-top (OTT) [1] services that provide access to content over the Internet, with an increasing number of OTT [1] companies offering these services. This has enabled users to enjoy unlimited internet video content irrespective of time or place. Despite this era, sports games are typically broadcast in real-time, which suggests that sports games are video content that is valuable to watch in real-time. The Sports DX Report [2], a summary of the Study Group on Rights for the Expansion of Sports Content and Data Business held jointly by Japan's Ministry of Economy, Trade and Industry and the Japan Sports Agency, states: "In an age where a variety of content such as TV dramas and movies can be viewed anytime, anywhere, live sports broadcasts are one of the few video contents that are worth watching in real-time. The one and only nature of sports content can lead to differentiation from competitors for business, which is thought to contribute to the increase in valuation." Given these factors, sports video content plays a crucial role in business.

However, sports games are typically quite long; the average length of a Major League Baseball (MLB) game during the 2022 season was 3 hours and 6 minutes [3]. Many users have difficulty watching games in real-time due to work, school, or time differences. Others have difficulty maintaining concentration for a long time or are interested only in specific

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

players. Moreover, unlike the other major sports, baseball games do not have a specific duration and are likely to run long. These factors have alienated younger fans from the game, presenting a serious problem. To address the issue, MLB has implemented several strategies to shorten games—introducing the batter's box rule [4] in 2015, the intentional walk [5] in 2017, limits to mound visits [6] in 2018, and the pitch timer [7] in 2023. Owing to the pitch timer [7], the average length of a game in the 2023 season was reduced to 2 hours, 42 minutes—a reduction of over 20 minutes. However, as baseball has no set game time, there are limits to the extent by which a game can be shortened.

In this paper, we focus on the MLB, one of the "Big Four," and propose a method of automatically classifying player at-bat results in real-time from baseball commentary, enabling fans to join in the enthusiasm for a game without watching it nonstop. The system also allows viewers to be updated on the course of a game quickly and efficiently compared to long-term viewing. In the current breaking news system, two to three people continuously watch the game, judging at-bat results and pitch types visually and inputting results manually; the goal of the proposed method in this paper is to support the current human-powered breaking news system. The contributions of our paper are as follows:

- We improve the real-time performance of traditional sports viewing support by using commentary.
- We propose a novel approach that combines pre-trained speech recognition and language models to keep track of the state of the game in real-time—not via social media or videos, but through sports commentary.
- We enhance the versatility and usability of existing speech recognition models by adapting them for real-time speech recognition, thereby broadening their range of potential applications.
- We create an adaptable, general-purpose system whose use of pre-trained models enables it to be applied to sports other than baseball.

The remainder of this paper is structured as follows: Section II discusses related work and its issues. Section III elaborates upon the proposed approach. Section IV evaluates the results, and finally, Section V concludes the paper.

II. RELATED WORK

To provide sports viewing support to users, a continuous understanding of the state of the game in real-time is necessary. Previous research uses X (formerly Twitter) to understand the state of the game in real-time [8], [9] instead of the approach proposed in this paper, which uses sports commentary. X displays posts in real-time and excels at the immediacy and reach of its information, so it is widely used as a social sensor not just in sports but in real-time event detection [10], such as stock prices [11], traffic reports [12], during earthquakes and other disasters [13]. We have also researched classifying at-bat results in baseball using real-time posts on X [14].

However, the majority of posts concern the posters' emotional responses, not detailed play information or at-bat results, making it challenging to derive at-bat results from these posts. Moreover, when multiple games are taking place simultaneously, posts about the games outside the target scope become noise, decreasing accuracy. Furthermore, it takes time for users to enter texts and make a post to X after a play, leading to reduced immediacy over commentary. X's limits on reading and posting and its introduction of paid features have been cited as the primary challenges in using X in this manner. These factors might accelerate the flight of users from X, making it difficult to obtain a continuous stream of posts from X in the future and posing problems for using the platform for at-bat result classification.

Research has also been conducted into generating highlight videos from game videos automatically [15], [16], [17], [18], [19], [20], as well as into scene classification [21], [22] and the generation of commentary [23], [24], [25]. However, numerous viewers prefer watching games in real-time, and highlight videos may not be sports viewing support. Furthermore, commentary generation requires listening to the game nonstop, so the time investment is no different from watching in real-time. In recent times, value has been identified in the commentaries themselves, with the uniqueness of commentaries becoming newsworthy; when machines become involved with commentaries, it can decrease excitement among fans.

The system proposed by this study uses language models and speech recognition models. After the debut of Generative Pre-trained Transformer (GPT) [26], a succession of language models pre-trained on large amounts of unlabeled text data, such as Bidirectional Encoder Representations from Transformers (BERT) [27] and Text-to-Text Transfer Transformer (T5) [28], have also appeared. These pre-trained models are called "Foundation models" [29] and frequently use the Transformer [30] architecture. Before the development of the Transformer, Natural Language Processing (NLP) used large amounts of labeled data geared to the task at hand, with the data used by the machine to train language understanding and category characteristics. However, as language understanding is standard and not task-specific, numerous pre-trained models using the Transformer, which excel at parallel processing, have been developed. These models have already learned general language patterns through pre-training, making it possible to build models that specialize in downstream tasks.

For speech recognition models, the significant cost of creating labeled audio data for pre-training has been cited as a challenge. Therefore, the debut of wav2vec 2.0 [31]—which takes self-supervised learning practices that have been successful in NLP and applies them to speech recognition—has dramatically enhanced speech recognition accuracy. NLP approaches for speech recognition based on the Transformer and BERT-based models have also been developed; for example, Conformer [32], which applies the Transformer and Convolutional Neural Networks (CNNs) in combination to speech recognition tasks, as well as HuBERT [33] and Google USM [34], which use both the Conformer and BERT pre-training methods. However, these models require fine-tuning and are not very versatile. Attention has therefore turned to Whisper [35]: an Automatic Speech Recognition (ASR) model that uses large-scale labeled audio data for pre-training and does not require fine-tuning. A speech recognition model trained with the large-scale Japanese audio corpus ReazonSpeech [36] is also being developed, with the development of supervised datasets also flourishing.

III. METHOD

Figure 1 shows the overall process of the proposed method. Its proposed method is divided into two parts: preprocessing and real-time processing. In preprocessing, speech recognition models are used to create labeled text data for fine-tuning; the pre-trained language models are then fine-tuned. In real-time processing, speech recognition models are used to produce transcripts of commentary from game videos automatically in real-time. The transcripts are input into the fine-tuned language models created during preprocessing, and at-bat results are classified. The data preparation, and fine-tuning processes conducted in preprocessing are explained in Sections A and B respectively, while real-time processing is explained in Section C.

A. DATA PREPARATION

For commentary in highlight videos, speech recognition models are used to produce transcripts automatically. The transcripts are then annotated. To classify at-bat results for this paper, it is necessary to prepare categories and corresponding labels for at-bat results in advance. Regarding the speech recognition models, due to the substantial cost associated with creating labeled audio data, we use not models that require fine-tuning like [31], [32], and [33] but Whisper and ReazonSpeech, both of which achieve accuracy comparable to [31], [32], and [33].

1) WHISPER

Whisper is an ASR model pre-trained on 681,070 hours of labeled audio data. Whisper does not require fine-tuning; it can perform various speech input tasks such as multilingual speech transcription, translation, spoken language identification, Voice Activity Detection (VAD), and alignment. Whisper architecture uses the Transformer, demonstrating robustness in large-scale supervised pre-training for speech recognition. Whisper offers 7,054 hours of Japanese labeled audio data. Five models with different numbers of parameters have been developed; we use the Large V2 [35] model, which features the largest number of parameters.

2) REAZONSPEECH

ReazonSpeech is a Japanese audio corpus composed of 15,735 hours of audio data. ASR models trained on ReazonSpeech have also been developed. The training uses a recipe provided by End-to-End Speech Processing Toolkit (ESPnet) [37]; the architecture is CTC segmentation [38]. Unlike Whisper, ReazonSpeech only produces transcripts from Japanese to Japanese; however, as ReazonSpeech features far more audio data than Whisper, it can produce accuracy equivalent to Whisper Large V2 using far fewer parameters.

B. FINE-TUNING

To specialize the pre-trained language models in classification tasks, fine-tuning is conducted using labeled text data created via the method described in Section III-A. The commentary transcripts include sports-specific expressions and keywords. For example, the commentary "He missed a hit." In such instances of commentary, classifying based solely on a word dictionary or keywords could misclassify the at-bat as a hit when it was out. To avoid misclassifications, comprehension of the transcript's meaning is essential. Therefore, in this paper, we use BERT and its derivations a distilled version of BERT (DistilBERT) [39], and A Robustly Optimized BERT Pretraining Approach (RoBERTa) [40]—which are capable of understanding context bidirectionally.

C. REAL-TIME PROCESSING

The proposed method of this paper, in order to keep track of the state of the game in real-time from sports commentary requires "real-time transcription of sports commentary" and "real-time classification of the sports commentary transcription." However, Whisper and ReazonSpeech accept only audio files as input. In other words, these models are limited to transcribing audio files. Therefore, it is impossible to recognize speech in real-time by using these models as they are. To address the limitations of existing models, we first capture the audio output from the speaker and use VAD to detect silent segments within the audio. When silence is detected, the preceding audio is saved temporarily as a file and is inputted into a speech recognition model for transcription. This process is looped to enable real-time transcription of sports commentary. Furthermore, by instantly inputting the transcription into a fine-tuned BERT model, we also make it possible to classify the at-bat results in real-time.

1) BERT

BERT is the first fine-tuning-based language representation model. BERT is distinguished by its capability to understand context not from left-to-right as in the past [26], but bidirectionally through masked language modeling in pre-training. Pre-trained on deep bidirectional representations, BERT can construct state-of-the-art models through fine-tuning using a small quantity of labeled text. Pre-trained models are dependent on the size of training data. Therefore, nowadays, many models that dwarf their predecessors in size have been developed. However, as the model size increases, so do deduction time, power consumption, and costs; the performance of PCs capable of running the model also becomes limited. Therefore, models derived from BERT were also

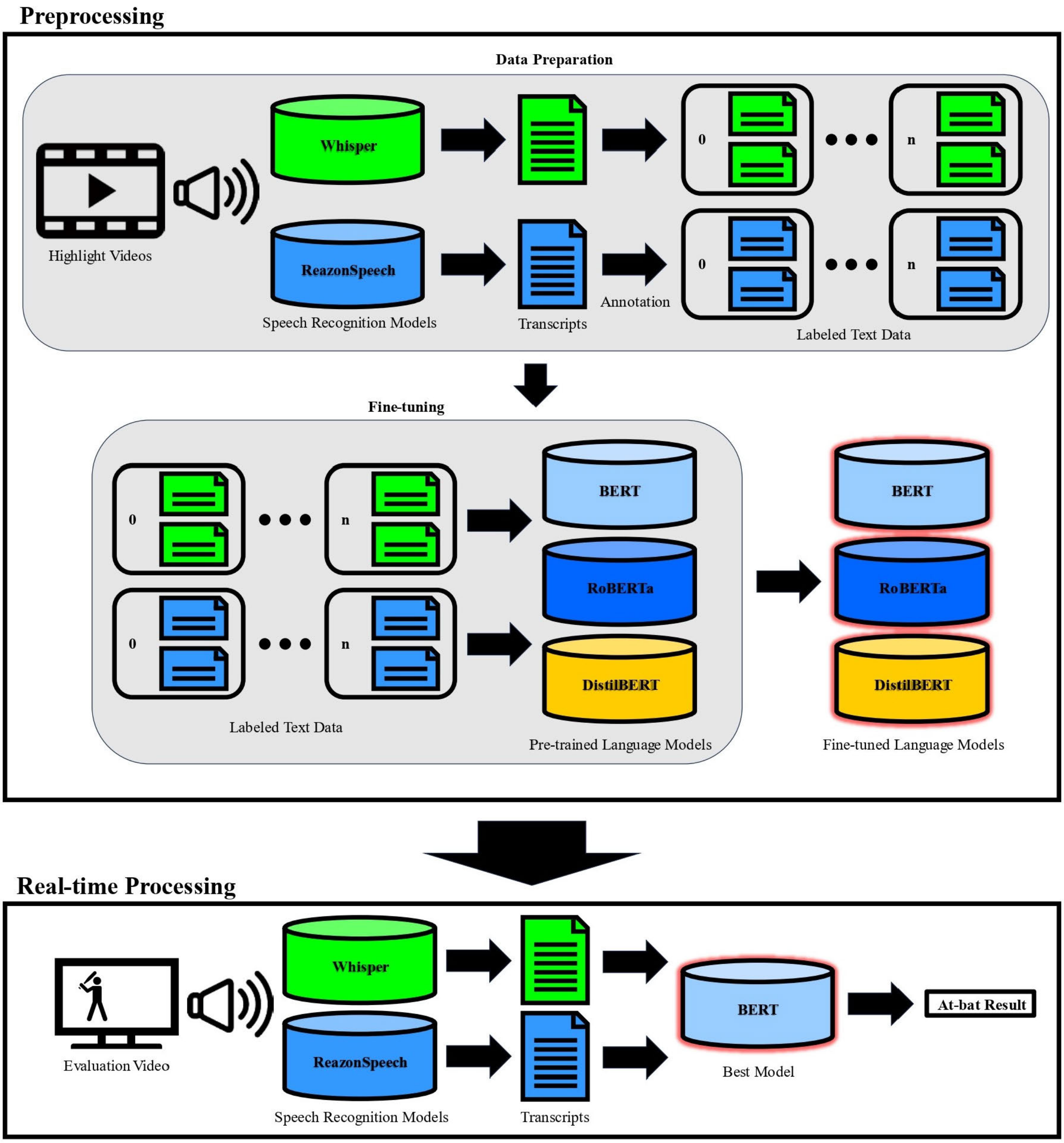

FIGURE 1. Overall process of the proposed method. The proposed method has two parts. We perform Data preparation as described in Section III-A and Fine-tuning as described in Section III-B. After that, we perform Real-time Processing as described in Section III-C.

developed—DistilBERT, which reduces the size of BERT while maintaining its performance in downstream tasks, and RoBERTa, which maintains the size of BERT but improves its performance in downstream tasks.

IV. EVALUATION

In this section, we use the method detailed in Section III to create labeled text data, fine-tune, and conduct processing in real-time, then evaluate the proposed method. The following hardware was used to conduct the fine-tuning and evaluations in this paper:

- CPU: Intel(R) Core (TM) i9-13900KF
- GPU: NVIDIA GeForce RTX 4090

A. DATASET

In this paper, we use speech from Japanese commentary highlight videos from the 2022 MLB season released on YouTube as a dataset for fine-tuning. A total of 134 videos

TABLE 1. Keywords for annotation.

Category	Keywords (Japanese)
BB	"shikyuu," "four-ball," "hit-by-pitch," "dead-ball"
Flyout	"left-fly," "center-fly," "right-fly," "left-eno," "center-eno," "right-eno"
Groundout	"goro"
Hit	"hit," "two-base," "three-base," "naiyaanda"
Home run	"home-run," "gou"
Strikeout	"sanshin"
Others	Contains any of the keywords but doesn't represent the at-bat results

were used. By focusing on the specific domain of MLB highlight video commentary and limiting the commentary to the Japanese, we can create a unique and innovative dataset that does not exist elsewhere. Furthermore, as the highlight video content features at-bat scenes compiled from the games, creating a dataset from highlight video content is significantly more efficient than creating one from the entire game footage. We use Whisper and ReazonSpeech for automatic transcription of the commentary audio in the highlight videos to generate transcripts for all video content. Table 1 lists the categories used in the paper. Transcripts generated using Whisper featured a lot of noise, and labeling the data using keywords was deemed difficult, so the data was labeled manually. By labeling manually, the quality of the dataset is ensured. On the other hand, transcripts generated using ReazonSpeech featured minimal noise, so this data was labeled using keywords. The keywords used for labeling are shown in Table 1. By using keywords for labeling, it is possible to reduce the time cost, which is the greatest challenge in data creation. Additionally, it becomes possible to eliminate annotator bias and construct a consistent dataset. Among the keywords, "left-eno," "center-eno," and "right-eno" were added with *eno*, indicating direction, because when only "left," "center," and "right" were used, home run commentary was mistakenly labeled as "Flyout." Also, "gou" was added for the number of home runs, as there were many commentaries that only mentioned the number of home runs, as there were many commentaries that only mentioned the number of home runs for the season, without including "home run." The categories are explained in Section B.

The evaluation used commentary from the March 12, 2023 (local date) WBC Pool B Japan vs. Australia game, the March 21, 2023 (local date) WBC United States vs. Japan championship game, and the July 11, 2023 (local date) MLB All-Star Game. Transcripts were generated in real-time and then input as necessary into the fine-tuned BERT model to classify at-bat results.

In this paper, we target speech in Japanese rather than English. Unlike English, Japanese has a large number of homophones and does not use spaces to separate words, requiring morphological analysis for contextual understanding. Therefore, speech recognition and contextual understanding are more complex tasks in Japanese than in English. There is also far less labeled speech and text data in Japanese than in English. In this paper, we targeted Japanese in the belief that a system demonstrating sufficient accuracy in Japanese would also be capable of being applied to other languages. Several BERT pre-trained on large-scale Japanese corpora have been released, but such corpora are extremely limited, with most of them based on Japanese Wikipedia. Therefore, out of the eight models we found, we selected the BERT BASE and BERT LARGE models with the same architecture as the original BERT [27], which are the latest versions released by Tohoku University.[1] For DistilBERT, out of the two models we found, only one was pre-trained on Japanese Wikipedia, so we selected the DistilBERT model released by Bandai Namco Research Inc.[2] For RoBERTa, since all seven models we found were pre-trained on Japanese Wikipedia, we chose the RoBERTa BASE model released by rinna Co., Ltd.,[3] which had the highest download count, and the RoBERTa LARGE model released by Waseda University.[4] We fine-tune the above five language models for at-bat classification and use them in our model.

B. CATEGORY

The Official Baseball Rules 2023 Edition [41] defines baseball as a game where the offensive team's objective is to have its batter become a runner and its runners advance, with the defensive team's objective is to prevent that. In other words, at-bat results are whether a batter becomes a runner or an out. There are three scenarios where a batter is out: a flyout, a groundout, or a strikeout. In other scenarios, the batter becomes a runner: a base on balls (BB), a hit, a hit by pitch, a home run, or an intentional walk. In this paper, we used seven annotated categories: "BB," "Flyout," "Groundout," "Hit," "Home Run," "Strikeout," and "Others." The highlight videos had very few BBs, hits by pitch, or intentional walks; therefore, we combined those three scenarios into the "BB" category. Likewise, the "Hit" category can be divided into single, double, and triple, but as triple was exceedingly rare, these three scenarios were combined into the "Hit" category. As for the "Others" category, if the transcript included characteristics of multiple categories—for example, "the pitcher gave up a hit before but is managing a strikeout"—classification under one of the established categories would reduce the Recall score; therefore, data examples that included target keywords but concerned the previous at-bat or discussed the previous day's scores were given the "Others" label.

A total of 840 examples were chosen and fine-tuned from the labeled text data—120 examples per category. Examples

[1] [Online]. Available: https://github.com/cl-tohoku/bert-japanese

[2] [Online]. Available: https://github.com/BandaiNamcoResearchInc/DistilBERT-base-jp

[3] [Online]. Available: https://github.com/rinnakk/japanese-pretrained-models

[4] [Online]. Available: https://huggingface.co/nlp-waseda

TABLE 2. Examples of sports commentary transcripts using ReazonSpeech (Japanese translated into English).

Category	Sports Commentary
BB	The pitch is inside, and it's a walk. Now we have runners on first and second.
	It's a dead ball. This has become a significant opportunity to get on base, with no outs and a runner on first.
	Shohei Ohtani's third at-bat was an intentional walk as first base was open.
Flyout	Overpowered by a 98-mile fastball, it's a shallow fly to the left. Three outs.
	It looked like a chance ball, but there's no extension to it, a flyout to center field.
	He hit the first pitch towards right but it's a flyout to Tucker in right field. Three outs.
Groundout	Groundout to first on an outside slider.
	Rengifo is a groundout to third. One out. Next up is the second batter, Shohei Ohtani.
	Induces a ground ball for a shortstop groundout. One out is recorded.
Hit	Good hit to the left. It's a wonderful hit.
	It's an extra-base hit, a double.
	Leadoff batter Nathaniel Lowe hits a triple. No outs with a runner on third.
Home run	It's gone. A super-fast bullet liner—it's 20th home run of the season for Ohtani, his first time in nine games.
	Trout's 28th home run.
	Back-to-back home runs. The Angels are pulling away.
Strikeout	The last pitch is a strikeout on an inside breaking ball.
	It's a called strikeout
	Ohtani was in the batter's box with a chance but struck out in the first inning.
Others	Today he was a base on ball, followed by an infield groundout, and then he got a hit.
	It was close to a home run.
	Leadoff batter Fletcher just missed a hit.

TABLE 3. Results and model comparisons from fine-tuning on labeled text data using Whisper (10 epochs, 30 epochs, 100 epochs, and 300 epochs).

Model	Time(s)	Loss	BB	Flyout	Groundout	Hit	Home run	Strikeout	Others
10 epochs									
BERT BASE	46.0	0.640	0.883	0.852	0.681	0.749	0.926	0.839	0.728
BERT LARGE	969.2	0.503	0.873	0.869	0.777	0.798	0.917	0.941	0.731
DistilBERT BASE	26.8	1.875	0.804	0.695	0.376	0.546	0.831	0.696	0.322
RoBERTa BASE	43.7	1.171	0.834	0.711	0.370	0.650	0.852	0.691	0.523
RoBERTa LARGE	972.4	0.645	0.794	0.738	0.745	0.763	0.917	0.858	0.647
30 epochs									
BERT BASE	109.0	0.533	0.861	0.876	0.790	0.785	0.944	0.893	0.759
BERT LARGE	3796.8	0.503	0.809	0.884	0.775	0.837	0.929	0.889	0.722
DistilBERT BASE	70.9	1.432	0.841	0.824	0.690	0.779	0.949	0.894	0.435
RoBERTa BASE	111.6	0.770	0.833	0.742	0.724	0.686	0.890	0.864	0.680
RoBERTa LARGE	2394.3	0.600	0.843	0.801	0.714	0.804	0.918	0.816	0.653
100 epochs									
BERT BASE	310.1	0.543	0.861	0.859	0.791	0.850	0.929	0.883	0.718
DistilBERT BASE	207.6	0.606	0.891	0.853	0.812	0.836	0.933	0.881	0.640
RoBERTa BASE	322.1	0.758	0.747	0.828	0.769	0.755	0.879	0.800	0.681
300 epochs									
DistilBERT BASE	521.3	0.590	0.886	0.855	0.854	0.816	0.923	0.895	0.686

of sports commentary transcripts used for fine-tuning, translated into English, are shown in Table 2.

C. FINE-TUNING AND COMPARISON USING WHISPER

Table 3 shows average values for the time needed for fine-tuning, loss, and classification accuracy using labeled text data composed of transcripts generated by using Whisper with each fine-tuned ten times for 10, 30, 100, and 300 epochs. However, when increasing the number of epochs resulted in a decrease in loss of at least 0.05, the number of epochs was further boosted, and fine-tuning was conducted. The labeled text data used during fine-tuning was randomly divided into training data (60%), validation data (20%), and test data (20%), then fine-tuned. In performance evaluation, fine-tuning was conducted ten times to reduce the validation in accuracy by category; average accuracy was then calculated, and the models were compared.

As shown in Table 3, fine-tuning with a small number of epochs produced lower classification accuracy for the ''Groundout'' and ''Others'' categories, which did not improve much even as the number of epochs increased. For the Groundouts, accuracy did not improve because the transcripts were longer than in the other categories and the expressions used had a great deal of variety, resulting in

TABLE 4. Results and model comparisons from fine-tuning on labeled text data using ReazonSpeech (10 epochs, 30 epochs, 100 epochs, and 300 epochs).

Model	Time(s)	Loss	BB	Flyout	Groundout	Hit	Home run	Strikeout	Others
					10 epochs				
BERT BASE	43.5	0.395	0.904	0.971	0.919	0.914	0.944	0.996	0.475
BERT LARGE	1370.3	0.335	0.930	0.958	0.987	0.936	0.903	0.974	0.575
DistilBERT BASE	25.0	1.839	0.870	0.896	0.813	0.867	0.934	0.996	0.056
RoBERTa BASE	46.7	0.977	0.806	0.791	0.754	0.828	0.943	0.932	0.183
RoBERTa LARGE	958.8	0.406	0.886	0.941	0.935	0.913	0.886	0.961	0.407
					30 epochs				
BERT BASE	108.4	0.346	0.923	0.975	0.939	0.963	0.947	0.973	0.609
BERT LARGE	3978.4	0.300	0.912	0.962	0.960	0.960	0.931	0.972	0.566
DistilBERT BASE	65.5	1.294	0.943	0.988	0.973	0.955	1.000	1.000	0.079
RoBERTa BASE	115.5	0.494	0.883	0.925	0.934	0.858	0.882	0.942	0.560
RoBERTa LARGE	3589.5	0.407	0.937	0.929	0.970	0.907	0.888	0.943	0.389
					100 epochs				
DistilBERT BASE	209.0	0.401	0.930	0.989	0.978	0.935	0.932	0.989	0.426
RoBERTa BASE	321.0	0.484	0.906	0.959	0.918	0.870	0.881	0.955	0.457
					300 epochs				
DistilBERT BASE	505.2	0.432	0.949	0.988	1.000	0.937	0.946	0.979	0.430

frequent speech recognition errors. For the Others, accuracy did not improve because the category featured vocabulary found in the other categories and similar transcripts. Furthermore, the impact of model size differences on accuracy is minimal, and it is believed that even the BASE model can sufficiently understand the context of the sports commentary.

The model used in the evaluation is DistilBERT BASE, which demonstrated the highest classification accuracy overall, fine-tuned at 300 epochs.

D. FINE-TUNING AND COMPARISON USING REAZONSPEECH

Table 4 shows the results of fine-tuning conducted using the same method detailed in Section III-C on labeled text data composed of transcripts generated by using ReazonSpeech. Each model proved capable of achieving accuracy close to 0.9 in all categories except Others. As with Section III-C, the "Others" category produced decreased accuracy due to the presence of vocabulary shared by the other categories and similar transcripts. Furthermore, as in Section IV-C, the impact of model size differences on accuracy is minimal, so it is believed that even the BASE model can sufficiently understand the context.

The model used in the evaluation was BERT BASE, which demonstrated the overall highest classification accuracy, fine-tuned at 30 epochs.

E. REAL-TIME PROCESSING RESULTS

In this Section, we evaluate models on the three games described in Section IV-A by using combinations of Whisper and DistilBERT BASE, as well as ReazonSpeech and BERT BASE, respectively.

TABLE 5. Precision, Recall, F-measure, and Support in Japan vs. Australia in the 2023 WBC.

Category	Precision	Recall	F-measure	Support
		Whisper and DistilBERT BASE		
BB	0.56	1.00	0.71	10
Flyout	1.00	0.65	0.79	17
Groundout	0.50	0.58	0.54	12
Hit	0.75	0.69	0.72	13
Home run	0.00	0.00	0.00	2
Strikeout	1.00	0.83	0.90	23
		ReazonSpeech and BERT BASE		
BB	0.71	1.00	0.83	10
Flyout	0.72	0.76	0.74	17
Groundout	0.55	0.50	0.52	12
Hit	0.88	0.54	0.67	13
Home run	1.00	0.50	0.67	2
Strikeout	0.96	1.00	0.98	23

1) JAPAN VS. AUSTRALIA IN THE 2023 WBC

Table 5 shows the Precision, Recall, F-measure, and Support of inputting transcriptions of the commentary of Japan vs Australia in the 2023 WBC into a fine-tuned best model. The Support is the number of occurrences of each category in the "accurate response" category.

It is evident that when Whisper and DistilBERT BASE are used, the "BB" and "Flyout" categories demonstrate a greater gap between Precision and Recall, with misclassifications into the "BB" category and with flyouts overlooked. Content tends to be misclassified as BB for inaccurate transcripts—particularly the ones that are ungrammatical or not well-formed. As for overlooked flyouts, the lack of a Line

TABLE 6. Precision, Recall, F-measure, and Support in United States vs. Japan in the 2023 WBC.

Category	Precision	Recall	F-measure	Support
Whisper and DistilBERT BASE				
BB	0.36	0.67	0.47	12
Flyout	0.70	0.80	0.74	20
Groundout	0.33	0.25	0.29	12
Hit	1.00	0.30	0.46	10
Home run	1.00	0.25	0.40	4
Strikeout	0.67	0.40	0.50	15
ReazonSpeech and BERT BASE				
BB	0.38	0.67	0.48	12
Flyout	0.70	0.70	0.70	20
Groundout	0.89	0.67	0.76	12
Hit	0.67	0.40	0.50	10
Home run	1.00	0.75	0.86	4
Strikeout	1.00	0.47	0.64	15

TABLE 7. Precision, Recall, F-measure, and Support in the 2023 MLB all-star game.

Category	Precision	Recall	F-measure	Support
Whisper and DistilBERT BASE				
BB	0.40	0.80	0.53	5
Flyout	0.77	0.62	0.69	16
Groundout	0.57	0.36	0.44	11
Hit	0.86	0.86	0.86	14
Home run	1.00	1.00	1.00	2
Strikeout	0.87	0.81	0.84	16
ReazonSpeech and BERT BASE				
BB	0.71	1.00	0.83	5
Flyout	0.88	0.94	0.91	16
Groundout	0.91	0.91	0.91	11
Hit	1.00	0.86	0.92	14
Home run	1.00	1.00	1.00	2
Strikeout	1.00	0.81	0.90	16

Drive category caused almost all line drives to be overlooked and misclassified as Groundout. All home runs were overlooked. As home runs result in a great deal of excitement, the crowd's roar would often drown out the commentators, leading to difficulties in transcription. For strikeouts, the transcripts include the word *Sanshin* (“strikeout”), resulting in favorable accuracy.

When ReazonSpeech and BERT BASE were used, the difference between precision and recall narrowed for the “BB” and “Flyout” categories; in particular, the improvement in recall for the “Flyout” category demonstrates the good performance of the models. As line drives were misclassified under the “Groundout” category, accuracy there did not improve. For home runs, ReazonSpeech accurately transcribed the “home run” that Whisper did not, resulting in improved accuracy.

Transcription using Whisper accuracy proved poor for speech recognition, with many misclassified transcripts being poorly formed grammatically or incomplete. On the other hand, ReazonSpeech demonstrated good transcription accuracy, with grammatically correct sentences; however, for short speech samples, the model could not recognize the first half of the sample, leading to misclassifications.

2) UNITED STATES VS. JAPAN IN THE 2023 WBC

Table 6 shows the Precision, Recall, F-measure, and Support of inputting transcriptions of the commentary of United States vs. Japan in the 2023 WBC into a fine-tuned best model.

When Whisper and DistilBERT BASE were used, the results demonstrated overall poor accuracy. In particular, there is a significant gap between precision and recall for the “Hit” and “Home run” categories, with overlooks common in both. Also, accuracy was very poor for the “Groundout” category; when accurately classified transcripts were compared with misclassified transcripts, the misclassified transcripts were poorly-formed grammatically, and there were difficulties in understanding the context, suggesting issues with the speech recognition, not the fine-tuning. As with the Japan vs. Australia game, when audience cheers and commentator voices overlapped, the model generated poor transcriptions and decreased accuracy.

When ReazonSpeech and BERT BASE were used, the results demonstrated improved accuracy over Whisper and DistilBERT BASE. The model performed strongly, able to classify line drives as flyouts half the time. However, the “BB” category has many misclassifications, with short speech samples and failure to transcribe the first half of the speech samples to blame. For the “Hit” category, Japanese uses a number of phrases other than “hit” to indicate hits, such as *sentaa mae* (“In front of center”) and *sentaa no mae ni ochiru* (“Falls in front of center”), and so hits are frequently misclassified as flyouts; there is, therefore, a need to re-examine the labeled text data and increase the amount used in fine-tuning. Similarly, for the “Strikeout” category, the shortness of the speech samples and the use of phrases other than “strikeout,” such as *subarashii koosu ni nagekonde kimashita* (“Pitched in a great course”) and *saigo wa Ohtani Shohei ga shimekukutta* (“Shohei Ohtani finishes it off.”), led many strikeouts to be misclassified as flyouts or BBs.

3) 2023 MLB ALL-STAR GAME

Table 7 shows the Precision, Recall, F-measure, and Support of inputting transcriptions of the commentary of the 2023 MLB All-Star Game into a fine-tuned best model.

The MLB All-Star Game included many scenarios where the commentator was not commenting on the game itself, such as segments where a baseball analyst was continuing to talk or interviewing a local player—with these segments representing 11 out of the 75 at-bats. As these segments do not communicate information about the state of the game, these were excluded from the evaluation.

TABLE 8. Average time for transcription and classification (Japan vs. Australia and United States vs. Japan in the 2023 WBC, and the 2023 MLB All-Star Game).

Model	Time(s)
Japan vs. Australia in the 2023 WBC	
Whisper [35] and DistilBERT BASE	0.543
ReazonSpeech [36] and BERT BASE	1.292
United States vs. Japan in the 2023 WBC	
Whisper [35] and DistilBERT BASE	0.537
ReazonSpeech [36] and BERT BASE	0.740
2023 MLB All-Star Game	
Whisper [35] and DistilBERT BASE	0.483
ReazonSpeech [36] and BERT BASE	1.190

Comparing the results of Tables 5 and 6 reveals that classification accuracy improved overall and that there were gaps between models in classification accuracy. Video of the All-Star Game reveals that compared to the WBC game, audience cheers were kept to a lower volume, and transcription accuracy was improved. Also, as the commentary team consisted of only one commentator and one baseball analyst, and as this was not a game played by Japanese teams, the commentator was more laid-back and less excited in their delivery, talking over each other less and facilitating, it is believed, a calmer broadcast and better transcripts.

For Whisper and DistilBERT BASE, precision in the BB category was low; as with the other games, low-accuracy transcripts tended to be classified under BB. In Japanese, "BB" is called *foabooru*("Four-ball"), so transcripts with the word *booru* ("ball") in them had a strong tendency to be misclassified as BBs. The dataset used for fine-tuning needs to be improved. For the "Groundout" category, recall was particularly low. This is due to the Japanese expression *saadogoro* ("grounder to third"), where *goro* is misrecognized as "ball" in the transcripts, leading to frequent misclassifications under BB.

ReazonSpeech and BERT BASE were confirmed a dramatic increase in accuracy. The results demonstrate that the proposed method in this paper can yield sufficient accuracy in quiet environments. The results of a "challenge" revealed some misclassifications in particular circumstances, such as scenarios where hits were misclassified as BBs.

4) AVERAGE TIME FOR TRANSCRIPTION AND CLASSIFICATION

Table 8 shows the average time it takes to detect a certain period of silence using VAD, transcribe audio, classify the transcript, and output the results. In this paper, real-time performance—the time required for transcription and classification—is important. As shown in Table 8, the proposed method in this paper shows a degree of immediacy that is difficult for the studies using the previous X-based method and for manual, human-powered systems to achieve manually.

V. CONCLUSION

In this paper, we proposed a method combining pre-trained speech recognition and language models to classify at-bat results in real-time from sports commentary. It also implemented fine-tuning to specialize BERT for the at-bat results classification task. The results confirmed that in situations where the impact of speech from parties other than commentators was low, the proposed method in this paper demonstrated sufficient performance in its task. In particular, for real-time performance, we performed at-bat results classification at a speed that is difficult to achieve manually.

Regarding the speech recognition models, we have demonstrated that the constraints of Whisper and ReazonSpeech have been eliminated, making them adaptable for real-time speech recognition. This enhances the models' versatility and usability, contributing to the expansion of their applications. Furthermore, by conducting a comparative evaluation between the two speech recognition models, Whisper and ReazonSpeech, we have identified the specific challenges of each model, providing insights for model selection.

In terms of data collection and data creation, by using highlight videos instead of full-game footage and limiting the language to Japanese, we have been able to efficiently construct a novel dataset with high originality that did not exist previously. We have also clarified the challenges the speech recognition models faced with the commentary, recommending manual labeling for transcripts by Whisper, and keyword-based labeling for transcripts by ReazonSpeech, thus ensuring the quality of each dataset. Additionally, by providing the keywords used for labeling, we believe this will serve as a guideline for constructing larger datasets in the future.

Regarding language models, for the task of classifying at-bat results from sports commentary, we found that even the smaller BASE model can be sufficiently used. Since we conducted comparisons across multiple language models, this outcome indicates the high quality of the fine-tuning dataset we have constructed, while for the speech recognition model, the difference in the size of the training data is reflected in the performance difference, although there are differences in the architecture.

As this paper identified a need for a Line Drive category, an increased amount of labeled text data, and more-accurate speech recognition in the proposed method, we would in the future like to increase the number of datasets and fine-tune the speech recognition models. Furthermore, there are currently seven categories; but, as Flyout and Groundout can be divided into more detailed at-bat results depending on the player position being processed and the Hit category has three variants (single, double, and triple), we would like to implement more detailed categories for at-bat results. Furthermore, we evaluate the proposed method targeting Japanese speech; we would also like to examine the method's applicability to other languages.

REFERENCES

[1] *Results of the Fact-finding Exercise on the Over-the-Top Programming Services*, Can. Radio-Telev. Telecommun. Commission, Ottawa, ON, Canada, Oct. 2011.

[2] *Sports DX Report*, Ministry Economy, Trade Ind. Jpn. Sports Agency, Chiyoda, Japan, Dec. 2022.

[3] C. Gough. (2000). *Average Length of Major League Baseball Games From 2000 To 2023*. Statista. Accessed: Oct. 30, 2023. [Online]. Available: https://www.statista.com/statistics/1310998/mlb-game-length/

[4] MLB Advanced Media. *Pace of Play*. Accessed: Oct. 30, 2023. [Online]. Available: https://www.mlb.com/glossary/rules/pace-of-play

[5] MLB Advanced Media. *Intentional Walk (IBB)*. Accessed: Oct. 30, 2023. [Online]. Available: https://www.mlb.com/glossary/standard-stats/intentional-walk

[6] MLB Advanced Media. *MLB Announces Pace of Play Initiatives for '18*. Accessed: Oct. 30, 2023. [Online]. Available: https://www.mlb.com/news/mlb-announces-pace-of-play-initiatives-c266718664

[7] MLB Advanced Media. (2023). *Pitch Timer (2023 Rule Change)*. Accessed: Oct. 30, 2023. [Online]. Available: https://www.mlb.com/glossary/rules/pitch-timer

[8] J. Kannan, A. R.-M. Shanavas, and S. Swaminathan, "Twitter sports: Real time detection of key events from sports tweets," *Trans. Eng. Comput. Sci.*, vol. 5, no. 6, pp. 36–60, Nov. 2017, doi: 10.14738/tmlai.56.3729.

[9] N. FUJIMOTO and T. USHIAMA, "Situational estimation of sports broadcasting using a character level auto-encoder for live tweets," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol.*, Dec. 2020, pp. 316–322.

[10] M. Hasan, M. A. Orgun, and R. Schwitter, "A survey on real-time event detection from the Twitter data stream," *J. Inf. Sci.*, vol. 44, no. 4, pp. 443–463, Aug. 2018, doi: 10.1177/0165551517698564.

[11] T. Mankar, T. Hotchandani, M. Madhwani, A. Chidrawar, and C. S. Lifna, "Stock market prediction based on social sentiments using machine learning," in *Proc. Int. Conf. Smart City Emerg. Technol. (ICSCET)*, Mumbai, India, Jan. 2018, pp. 1–3.

[12] G. A. Neruda and E. Winarko, "Traffic event detection from Twitter using a combination of CNN and BERT," in *Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS)*, Depok, Indonesia, Oct. 2021, pp. 1–7.

[13] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: Real-time event detection by social sensors," in *Proc. 19th Int. Conf. World Wide Web*, Raleigh, NC, USA, Apr. 2010, pp. 851–860.

[14] K. Sakamoto, R. Ikeda, and Y. Ueda, "Classification of texts describing baseball batting results on Twitter using BERT," *Studies Sci. Technol.*, vol. 12, no. 1, pp. 93–99, Jun. 2023, doi: 10.11425/sst.12.93.

[15] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proc. 8th ACM Int. Conf. Multimedia*, Marina del Rey, CA, USA, Oct. 2000, pp. 105–115.

[16] C.-C. Cheng and C.-T. Hsu, "Fusion of audio and motion information on HMM-based highlight extraction for baseball games," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 585–599, Jun. 2006, doi: 10.1109/TMM.2006.870726.

[17] C. Xu, Y.-F. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang, "Using webcast text for semantic event detection in broadcast sports video," *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1342–1355, Nov. 2008, doi: 10.1109/TMM.2008.2004912.

[18] Y. Lee, H. Jung, C. Yang, and J. Lee, "Highlight-video generation system for baseball games," in *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, Nov. 2020, pp. 1–4.

[19] K. Arasawa and S. Hattori, "Automatic baseball video tagging based on voice pattern prioritization and recursive model localization," *J. Adv. Comput. Intell. Intell. Informat.*, vol. 21, no. 7, pp. 1262–1279, Nov. 2017, doi: 10.20965/jaciii.2017.p1262.

[20] K. Hirasawa, K. Maeda, T. Ogawa, and M. Haseyama, "Detection of important scenes in baseball videos via bidirectional time lag aware deep multiset canonical correlation analysis," *IEEE Access*, vol. 9, pp. 84971–84981, 2021, doi: 10.1109/ACCESS.2021.3088284.

[21] W. Hua, M. Han, and Y. Gong, "Baseball scene classification using multimedia features," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Lausanne, Switzerland, Aug. 2002, pp. 821–824.

[22] R. Ando, K. Shinoda, S. Furui, and T. Mochizuki, "A robust scene recognition system for baseball broadcast using data-driven approach," in *Proc. 6th ACM Int. Conf. Image video Retr.*, Amsterdam, The Netherlands, Jul. 2007, pp. 186–193.

[23] T. Kumano, M. Ichiki, K. Kurihara, H. Kaneko, T. Komori, T. Shimizu, N. Seiyama, A. Imai, H. Sumiyoshi, and T. Takagi, "Generation of automated sports commentary from live sports data," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jeju, South Korea, Jun. 2019, pp. 1–4.

[24] B.-J. Kim and Y.-S. Choi, "Automatic baseball commentary generation using deep learning," in *Proc. 35th Ann. ACM Symp. Appl. Comput.*, Brno, Czech Republic, Mar. 2020, pp. 1056–1065.

[25] H. Mkhallati, A. Cioppa, S. Giancola, B. Ghanem, and M. Van Droogenbroeck, "SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Vancouver, BC, Canada, Jun. 2023, pp. 5074–5085.

[26] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (2018). *Improving Language Understanding By Generative Pre-Training*. Accessed: Oct. 30, 2023. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

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

[28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[29] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.-N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[31] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2020, pp. 12449–12460.

[32] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. InterSpeech*, Shanghai, China, Oct. 2020, pp. 5036–5040.

[33] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3451–3460, 2021, doi: 10.1109/TASLP.2021.3122291.

[34] Y. Zhang et al., "Google USM: Scaling automatic speech recognition beyond 100 languages," 2023, *arXiv:2303.01037*.

[35] A. Radford, J.-W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Mach. Learn.*, Honolulu, HI, USA, 2023, pp. 28492–28518.

[36] Y. Yin, D. Mori, and S. Fujimoto, "ReazonSpeech: A free and massive corpus for Japanese ASR," in *Proc. 29th Ann. Meeting Assoc. Natural Lang. Process.*, Okinawa, Japan, 2023, pp. 1134–1139.

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

[38] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, "CTC-segmentation of large corpora for German end-to-end speech recognition," in *Proc. Int. Conf. Speech Comput.*, St. Peterburg, Russia, Oct. 2020, pp. 267–278.

[39] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," in *Proc. 5th Workshop Energy Efficient Mach. Learn. Cognit. Comput. Co-Located Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2019, pp. 1–5.

[40] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[41] *Official Baseball Rules 2023 Edition*, 2nd ed., Office Commissioner Baseball, New York, NY, USA, Mar. 2023.

RIKU IKEDA (Graduate Student Member, IEEE) was born in Kanazawa, Ishikawa, Japan, in 1999. He received the B.S. degree in production systems engineering and sciences from Komatsu University, Japan, in 2022, where he is currently pursuing the M.S. degree with the Graduate School of Sustainable Systems Science. His research interests include web mining, natural language processing, and speech processing.

KAZUMA SAKAMOTO was born in Kyoto, Japan, in 1991. He received the B.S., M.S., and Ph.D. degrees in informatics from Kansai University, Japan, in 2015, 2018, and 2021, respectively.

Since 2021, he has been an Assistant Professor with the Faculty of Production Systems Engineering and Sciences, Komatsu University, Japan. His research interests include web mining, natural language processing, image processing, and sports informatics.

YOSHIHIRO UEDA was born in Kanazawa, Ishikawa, Japan, in 1960. He received the B.S. and M.S. degrees in administration engineering from Keio University, Japan, in 1981 and 1983, respectively, and the Ph.D. degree in mathematical information science from Kanazawa University, Japan, in 2001.

From 1989 to 2021, he was a Researcher/ Research Manager with the Industrial Research Institute, Ishikawa. Since 2021, he has been a Professor with the Faculty of Production Systems Engineering and Sciences, Komatsu University, Japan. He is the author of more than 30 articles and more than nine inventions. His research interests include productivity improvement, labor saving, automation using data science, and human centric system development.