THE JOURNAL ON TECHNOLOGY AND PERSONS WITH DISABILITIES

# A Timing Determination Method for the Insertion of Automated Audio Descriptions into Live TV Sports Programs

Manon Ichiki, Hiroyuki Kaneko
NHK Science and Technology Research Laboratories
Atsushi Imai, Tohru Takagi
NHK Engineering Systems
ichiki.m-fq@nhk.or.jp, kaneko.h-dk@nhk.or.jp, imai.atsushi@nes.or.jp, takagi@nes.or.jp

## Abstract

We are conducting research on automated audio descriptions (AADs) to help visually impaired people enjoy live TV programs. This technology uses real-time data (who, when, what is done, etc.) that has been automatically or manually generated from a sports match. The data are turned into audio descriptions by a voice synthesizer, and these are then distributed simultaneously with the broadcast audio. However, a problem occurs when AADs overlap with the live commentary, as the listener is put in the difficult position of having to listen to both the commentaries and the AADs at the same time. Therefore, overlaps need to be prevented in order for the sports program to be understood. In this work, we propose a timing determination method to insert AADs into live sports programs. The proposed method predicts the end of each utterance in a commentary, and AADs are then inserted after the commentaries have finished. In this method, the difference between the long- and short-term moving average of the fundamental frequency (F0) is utilized to predict the end of utterances. Visually impaired people evaluated the ease of listening to both commentaries and AADs and indicated that our method makes it easier for them to listen.

## Keywords

Audio Description, Broadcasting, Visually Impaired People, Fundamental Frequency (F0)

## Introduction

Audio descriptions are mainly used for the benefit of visually impaired people as a supplementary service to provide information that cannot be obtained from the main audio alone. At present, the service is recorded in advance and mostly used for such programs as TV dramas. Audio descriptions are not currently provided on live broadcasts, as it is extremely difficult for human operators to provide, in real time, comparable descriptions with appropriate timings. In response to this difficulty, we have been studying an automatic system for generating and inserting audio descriptions into live sports relay programs.

We have already developed techniques for the automatic, real-time generation of play-by-play descriptions for use on sports relays with stadium noise only and no play-by-play commentary (Taro et al.) (Kiyoshi et al.). These descriptions can be input and broadcast sequentially from the venue in real time. The system receives technical competition data feeds on such items as points and offences, generates sentences to describe what has just happened, and converts these into verbal statements by means of voice synthesis. We have used this system to provide fully automatic play-by-play audio on live sports relays streamed on the Internet. This technology is positioned as the first step towards the achievement of automated audio descriptions (AADs). Our next step is to select such details as the name of the event and the names of competitors (i.e., the who, when, and what) superimposed in text on the screen, which are liable to be precluded in live commentaries, from the play-by-play descriptions and to provide them in the form of concise verbal expressions as AADs. This makes it possible for even those who have difficulty obtaining information from the image to get basic information about the game.

**Issue with Overlapping of Live Television Commentaries and AADs**

In automatic audio descriptions, where the broadcast audio also includes the voices of actual commentators, automatic audio description is liable to overlap. Such overlap, of course, makes it difficult for the listener to follow what is being said. We performed a preliminary study (Atsushi et al.) to investigate the incidence of overlap between commentary voices and AADs delivered as quickly as possible after getting the competition data. The results showed very different rates of overlap depending on the sport: 20% for tennis and 70% for basketball. However, in a five-grade evaluation on whether they would use the service, all respondents gave an evaluation value of 5 regardless of the amount of overlap. This suggests that AAD may be useful as a voice assistive information service. On the other hand, in order to make it easy to listen to the service in spite of the overlaps, all respondents mentioned the need to adjust the acoustical conditions between commentary voices and AADs.

We focused, therefore, on the following four investigations.

1. Ease of listening depending on the difference in voice loudness between commentary voices and AADs.
2. Effect of voice characteristics of the text-to-speech synthesizer on ease of listening to sounds the viewers want to listen to.
3. AAD output through a speaker placed in a different location from the TV speaker.
4. Presentation of AADs at a timing to prevent overlap with live commentaries.

We conducted the following basic investigations for 1–3 (Manon et al.). Among these, the third method is regarded as the most effective because 74% of the participants answered that both the live commentary and the AADs were easy to listen to. We also discuss a fourth investigation of a method to improve ease of listening in the next Chapter.

**A Method for Predicting the End of an Utterance**

In order to prevent the overlaps, we propose a strategy in which AADs are inserted after the detection of the end of an utterance during live commentary. Fig. 1 shows an example of this strategy. Three AADs in part (A) overlap with live commentary when they are inserted shortly after their generation. In part (B), however, the AADs are inserted after the end of every utterance so that the overlaps are prevented.

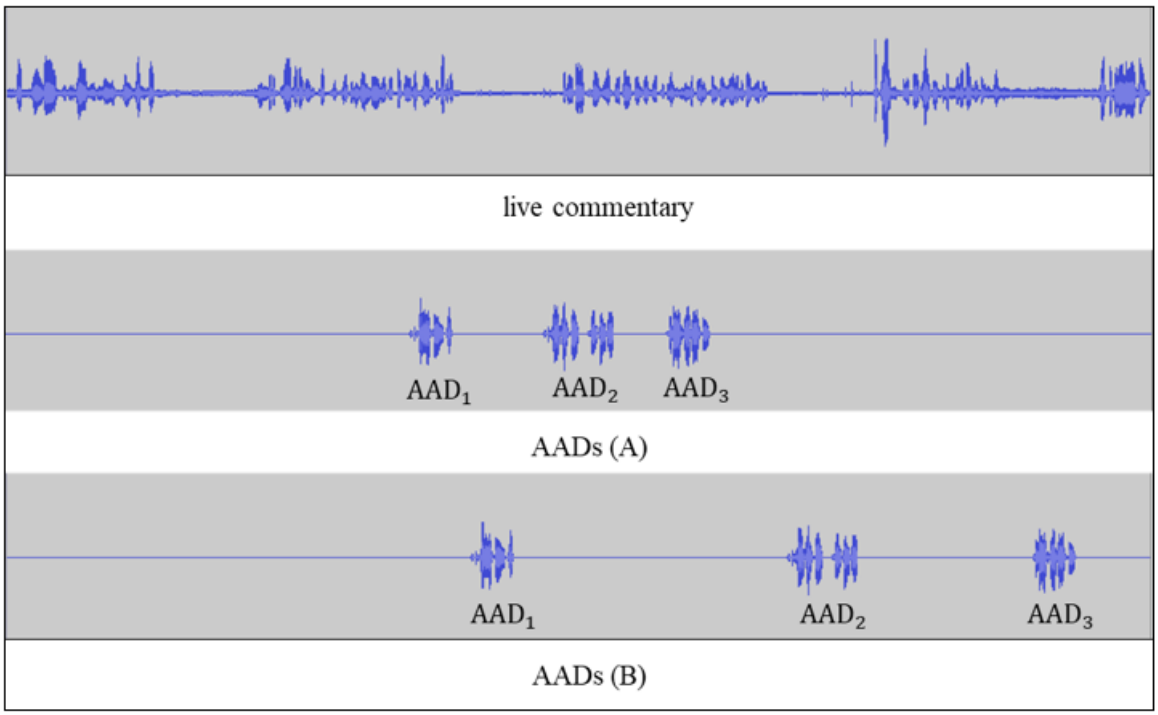


Fig. 1 Sound Waveforms of a Live Commentary and AADs Inserted with Two Techniques.

To grasp when an utterance is about to end, we focused on the acoustic features of the end of spontaneous utterances, namely, how the fundamental frequency (F0) tends to become lower from the start to the end of an utterance (Yuichi et al.). Slight fluctuations arise if the F0 data are used in their raw state, producing frequent lowering parts during utterances, so we can obtain the moving average (MA) values of the F0 to grasp the utterance's trends. MA values

were obtained retroactively at five-millisecond intervals (Tohru et al.) for the most recent few seconds to each present moment.

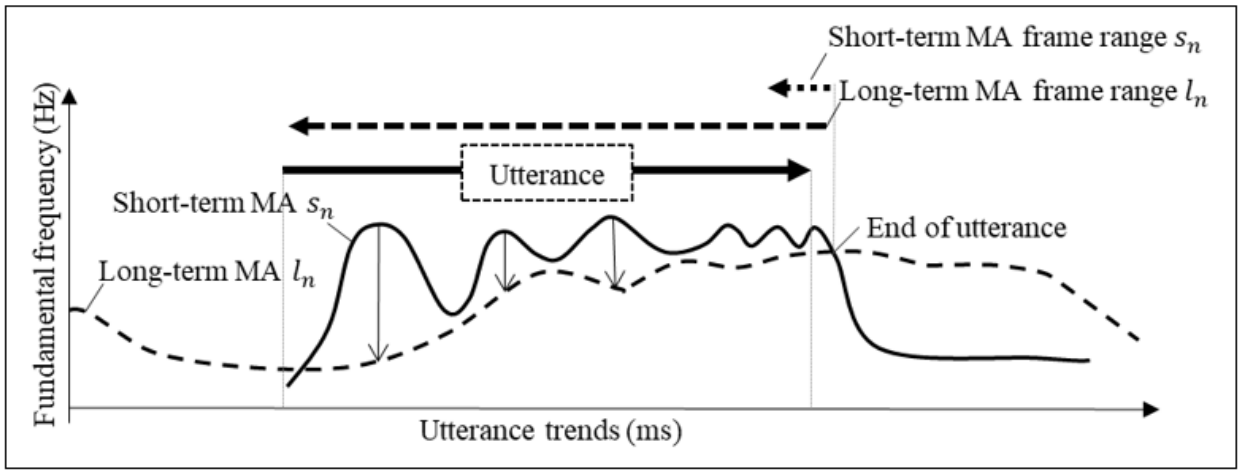


Fig. 2 Detecting End of Utterance by Moving Average (MA).

The value $c_n$ used to judge the end of an utterance was calculated as shown in Fig. 2 using two data items: $s_n$ for the short-term moving average (SMA) and $l_n$ for the long-term moving average (LMA). The SMA was calculated for a one-second period to grasp the F0 movement of short utterances. However, the SMA also catches several lowering parts in the F0 during these utterances, so it is insufficient to identify the end of the utterances. For this reason, we also use the LMA to predict when the utterances end. The LMA of the F0 is calculated for a four-second period, which is the typical length of a spoken utterance. When the value of the SMA for the F0 of the most recent utterance becomes lower than the MA of a somewhat longer term, this is taken as a sign that the utterance is coming to an end.

*Evaluation of the Timing Determined by Our Method*

We examined 19 programs featuring nine different sports (broadcasts of badminton, tennis, table tennis, etc.). A roughly three-minute section was selected from each program to make a combined total of 57 minutes of audio for the evaluation. The performance was evaluated

with respect to human judgment on when each utterance ended, allowing a margin of 800 milliseconds on both sides. A prediction by our method was counted as correct if it fell within this range.

*Evaluation Results*

The recall rate is the ratio of correct predictions to the total number of utterance endings indicated by human judgment. The precision is the ratio of correct predictions to the total number of predictions.

The overall recall rate was 77%, showing a fairly high level of correct prediction. Tennis and table tennis are sports in which the progression of the match is quite predictable, so the play-by-play announcers and commentators tend to wait a while for the cheering to subside before speaking about each point. This manner is similar to how our method works, and we presume this is why the recall rate improves. However, the precision was less than 40% for badminton and soccer, among others, so the performance was clearly not ideal. A prediction of nearly double the total number of endings indicated by human judgment indicates excessive selection.

**Discussion**

*Subjective Evaluations*

We evaluated the ease of listening when presenting live commentaries and AADs. The timing of each AAD was determined using our method. The experimental conditions are listed in Table 1. The presentation stimuli were a mix of live commentary and AADs. These presentation stimuli were made without the proposed method, called (A), and with the method, called (B). Eight different scenes of a women's table tennis game were selected to generate two-minute stimuli, with four scenes for (A) and four for (B). Three stimuli were randomly selected from the four every (A) and (B), and these stimuli were presented to a participant in random order. At this

time, nine of 18 participants listened to two stimuli from (A) and one stimulus from (B), and the other nine participants listened to the remaining one from (A) and the remaining two from (B). Therefore, the total number of trials for each of (A) and (B) was 27, and stimuli were presented a total of 54 times. As shown in Fig. 3, a participant answered with a five-point evaluation after listening to a stimulus sound for two minutes. Fig. 4 shows the placement of the speakers, sound localizations, and the listening position of the participant.

Table 1. Conditions of Ease-of-Listening Experiment.

| **Category** | **Condition** |
|---|---|
| Participants | 18 visually impaired people (average age: 50) |
| Evaluation stimulus | Three scenes of women's table tennis |
| Live commentary | Male (in Japanese) |
| Voice synthesizer for AAD | Female (in Japanese) |
| Presentation method | A (Without our method)<br>B (With our method) |
| Subjective assessment item | To what extent did you evaluate the ease of listening both to live television commentaries and AADs?<br>Five-grade evaluation (1–5) |

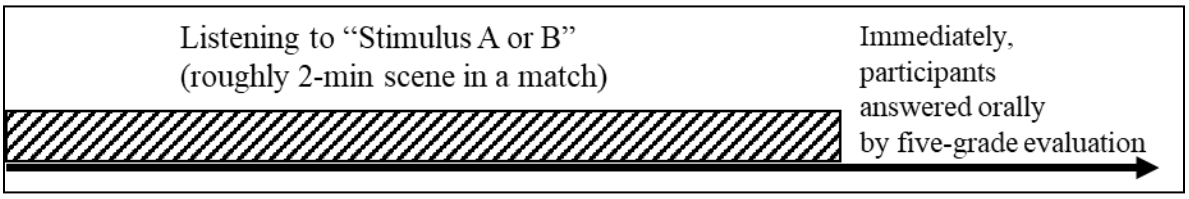


Fig. 3 Experiment with Stimulus Sounds and Subjective Assessment item.

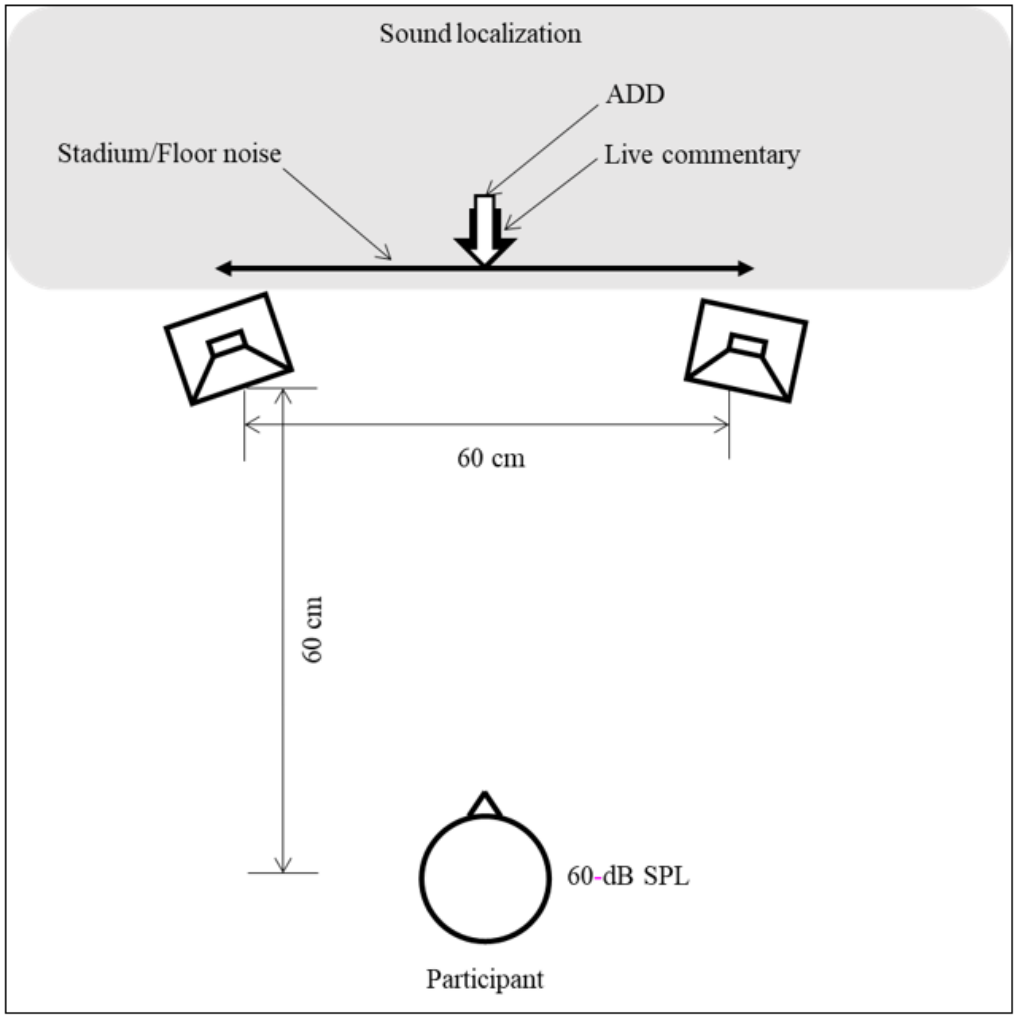


Fig. 4 Sound Presentation Conditions.

Table 2 shows the total and average results of the five-grade evaluations with (A) and (B), and Fig. 5 shows the corresponding bar graphs with 95% confidence intervals (CIs). The average value and standard deviation of (A) without our method was 3.48 ± 1.33 (2.15 to 4.81), and that with our method (B) was 4.56 ± 0.80 (3.76 to 5.36). The 95% CI for (A) was 3.48 ± 0.53 (2.95 to 4.01), and that for (B) was 4.56 ± 0.32 (4.24 to 4.88). Comparing these results, we can see that the lower limit value 4.24 of (B) had a difference of +0.23 from the upper limit value 4.01 of (A). This demonstrates that there is a significant difference in the critical region of 5%, as the two 95% CIs did not overlap. The results of this evaluation confirm that the proposed method

(B) is easy to listen to. After the experiment, we asked participants about their impressions. Their responses included “If AADs overlap with live commentary, I need to concentrate on both commentaries, so I had no idea which one I should listen to” and “It’s better for ease of listening that AADs don’t overlap with the live commentary”.

Table 2a. Total Evaluation Value with Each Scene Without Our Method (A).

| Scene | A1 | A2 | A3 | A4 | total | avg. |
|---|---|---|---|---|---|---|
| **Presentation time** | 6 | 8 | 6 | 7 | 27 | – |
| **Total value** | 22 | 29 | 21 | 22 | 94 | 3.48 |

Table 2b. Total Evaluation Value with Each Scene with Our Method (B).

| Scene | B1 | B2 | B3 | B4 | total | avg. |
|---|---|---|---|---|---|---|
| **Presentation time** | 8 | 7 | 6 | 6 | 27 | – |
| **Total value** | 36 | 35 | 26 | 26 | 123.00 | 4.56 |

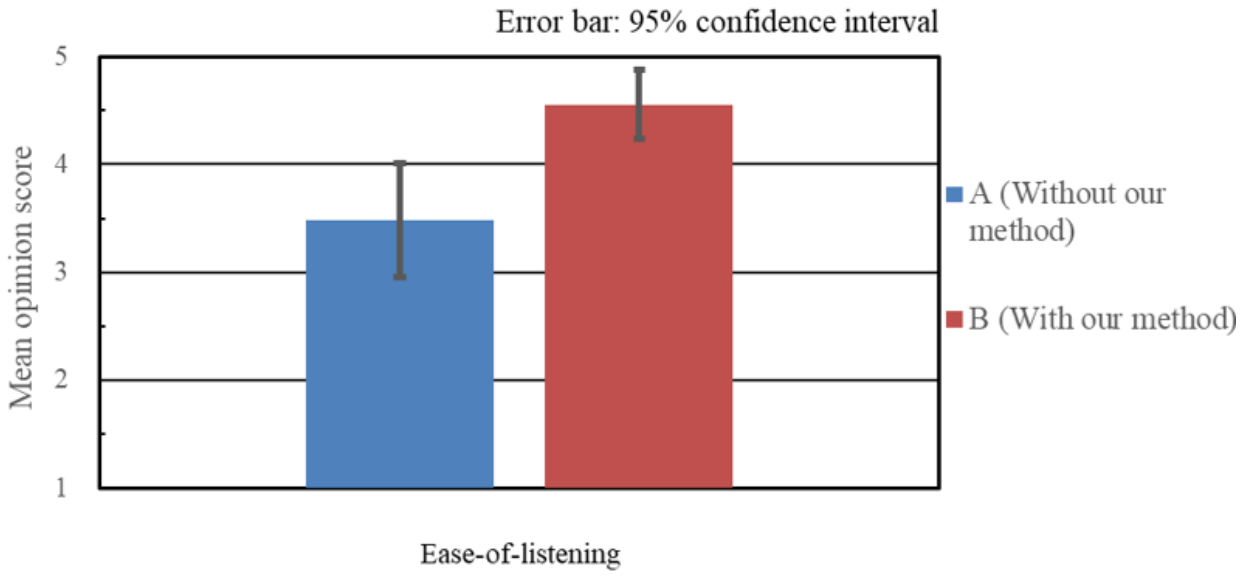


Fig. 5 Subjective Evaluation of Ease-of-Listening to Both Commentaries and AADs.

## Conclusion

We are developing an AAD system that will enable all viewers to enjoy sports programs regardless of any visual impairment. In this paper, we presented a timing determination method for the insertion of AADs into live TV commentary to solve the overlap problem, and we demonstrated its effectiveness through subjective evaluation.

As the next step, in order to improve the ease of listening to both live commentary and AADs, we will develop an application for smartphones that can present AADs at different locations from the TV speaker and proceed with providing this service via Internet distribution. This method can be used not only for live TV sports programs but also for other shows as well, demonstrating its wide range of applicability in the future.

**Works Cited**

Atsushi, I., et al. "Study of Automated Audio Description Service for Live TV Sports Program." *32nd CSUN Assistive Technology Conference*, California, 2017, ENT-014.

Kiyoshi, K., et al. "Automatic Generation of Audio Description for Sports Programs." *International Broadcasting Convention*, Amsterdam, 2017.

Manon, I., et al. "Study on Automated Audio Descriptions Overlapping Live Television Commentary." *16th International Conference on Computers Helping People with Special Needs*, Springer, Linz, 2018, pp. 220-224.

Taro, M., et al. "Automatic Generation of Audio Description for Olympics / Paralympics Programs." *NAB (National Association of Broadcasters) Show Conference, Broadcast Engineering and Information Technology Conference*, Las Vegas, 2017, N256.

Tohru, T., et al. "A Method for Pitch Extraction of Speech Signals Using Autocorrelation Functions through Multiple Window-Lengths." *IEICE*, 1997, pp. 1341–1350. (in Japanese)

Yuichi, I., et al. "Projectability of Transition-Relevance Places Using Prosodic Features in Japanese Spontaneous Conversation". *INTERSPEECH*, Italy, 2011, pp. 2061–2064.