AUTOMATIC GENERATION OF AUDIO DESCRIPTIONS FOR SPORTS PROGRAMS

Kiyoshi Kurihara[1], Atsushi Imai[1], Hideki Sumiyoshi[1], Yuko Yamanouchi[1], Nobumasa Seiyama[1], Toshihiro Shimizu[1], Shoei Sato[1], Ichiro Yamada[1], Tadashi Kumano[1], Reiko Tako[1], Taro Miyazaki[1], Manon Ichiki[1], Tohru Takagi[2], Susumu Oshima[1] and Koji Nishida[1]

[1]NHK, Japan and [2]NHK Engineering Systems, Japan

ABSTRACT

NHK has developed a means of automatically generating auxiliary audio descriptions from metadata for use in live TV sports programs. Audio description services are important for helping visually impaired persons enjoy TV programs, but such services are currently available for only a handful of programs because many studio resources and personnel are required to create audio descriptions, and it is especially difficult to produce such descriptions during live broadcasts. The method described in this paper has the potential to overcome these obstacles.

The system that we constructed for the Rio Olympic and Paralympic Games consists of commentary text generation and text-to-speech (TTS) processes. The commentary text generation process generates commentary appropriate to the situation for each piece of event data accepted by the system, and the TTS part converts it into natural speech.

We ran the system during the Rio Olympic and Paralympic Games, and it provided both caption and audio descriptions for over 2,000 sporting contests.

INTRODUCTION

We present an automatic system for live generation of audio descriptions for TV sports programs. Our research on automatic generation of audio descriptions has two main goals. One is to provide efficient and effective automatic program commentaries that are useful for sighted persons. The other is to provide audio descriptions that can help visually impaired persons get more out of TV.

Although they have been recognized as a helpful program service, audio descriptions are currently provided for only 10% (1) of the programs broadcast in Japan. In most cases, audio descriptions can only be attached to content during post-production in TV studios and control rooms, and the process entails the effort of sports announcers, directors, and technical staff. The expense, in terms of studio and personnel costs, and the difficulty of adding live audio descriptions to programming, has so far constrained the growth of audio-

description broadcasting. Automatic generation of audio descriptions, on the other hand, has the potential to solve these problems and improve the penetration of the service in broadcasting. In our development, we first constructed a prototype system for automatically generating audio descriptions from event data gathered at the Rio de Janeiro Olympic and Paralympic Games in 2016. Our aim was to produce test programs by automatically attaching commentary to the video footage, without manually adding any commentary, and examine the issues which arose.

Visually impaired persons as well as sighted persons stand to benefit from automatically generated audio descriptions attached to commentary-free international TV signals (2).

The programs broadcast during the Olympic and Paralympic Games only partly conformed to our program schedules. In Japan, international TV signals of events which could not be broadcast have been distributed domestically for exclusive viewing in Japan via the Internet. The use of automatically generated audio descriptions for such non-broadcasted events can help all viewers enjoy this kind of coverage.

OVERVIEW OF AUTOMATICALLY GENERATED AUDIO DESCRIPTIONS

Figure 1 is an overview of the automatic system for generating audio descriptions. The system generates sports commentary sentences from sports metadata and then reads them out loud by text-to-speech synthesis (TTS). Audio descriptions can be automatically generated from the sports metadata and video footage. The descriptions are then embedded in a sports program to which no prior commentary had been attached. The descriptions are generated immediately when the sports metadata are received and can, therefore, also be used live. We developed our automatic generation system for Olympic Data Feed (ODF) (3)(4), i.e. sports metadata, and ran demonstrations at the Rio Olympic and Paralympic Games. The ODF and IPVandA (IP Video and Audio) (3) were produced at each venue and the International Broadcast Centre (IBC) and distributed to NHK. The international TV signals were sent by IPVandA in MPEG-2 format over 45 IP transmission

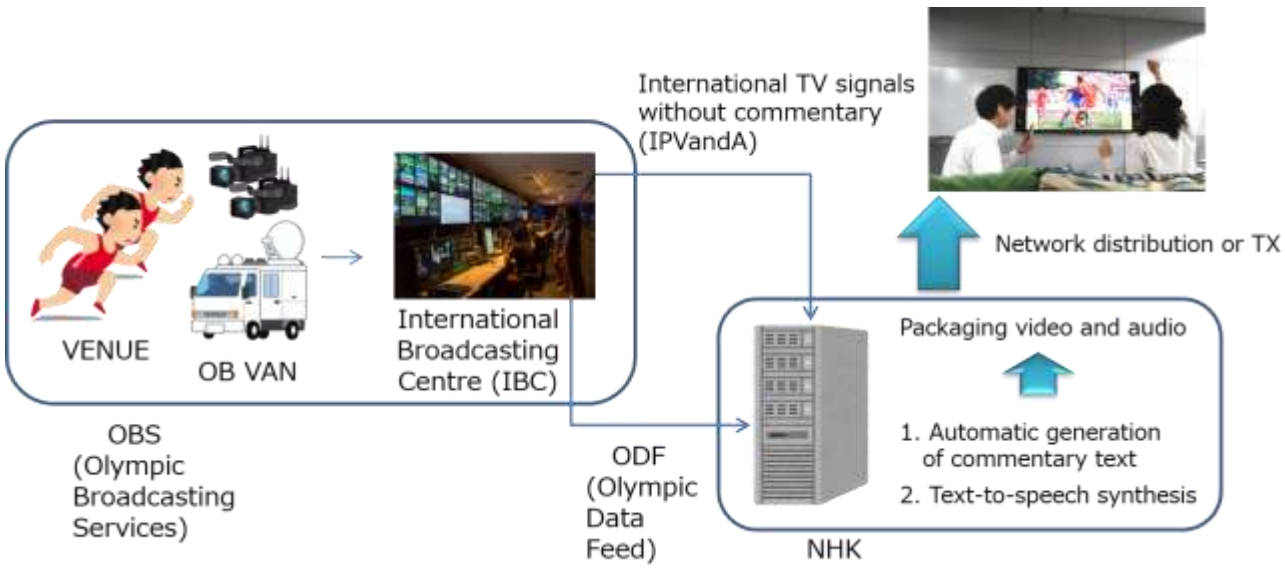

Figure 1 – Overview of automatic generation of audio description for sports programs

channels. Video shot at the venue was sent to the IBC via an OB van, and ODF staff at the venue and at the IBC monitored these programs. The ODF and IPVandA were distributed to NHK's server to trigger the automatic generation of the audio descriptions. The commentary text generation and TTS processes automatically produced the audio sports commentary. The generated audio descriptions were packaged with the international TV signals; no manual commentary was attached during production of these sports programs.

INTERNATIONAL TV SIGNALS AND OLYMPIC DATA FEED (ODF)

We developed a new method for automatic producing sports programs that uses the ODF and international TV signals received without commentary.

The international TV signals are provided as video footage for rights holding broadcasters of each country to use in their own domestic broadcasts. The international TV signals are produced fairly and neutrally with no bias towards any country's athletes or team, and normally include no commentary and only the bare minimum of visual information (2). TV stations in each country then edit the footage into their own unique video package and add their own commentary. This makes it possible to produce unique programs with relatively little labour. NHK distributes the international TV signals for exclusive domestic viewing in Japan on the Internet at the same time as the broadcasts. However, in view of the vast amount of video footage, it is impossible to provide unique commentaries and edits for all of them and they have to be distributed in their raw form without editing. It is also difficult for viewers to understand the state of a competition from the raw video alone; making audio descriptions helps not only visually impaired people, but also sighted persons, to comprehend and enjoy the programming. We considered that the generation of audio descriptions from the ODF would be way to overcome the associated problems.

The ODF is official, real-time sports metadata prepared in XML format and distributed on the Internet by the Olympic Broadcasting Services (OBS). The data consists of athlete-related information, including the competitors' names, start lists, and current records such as scores, play-by-play event descriptions, and statistical data (3). The ODF is often used to superimpose current scores on the screen. Rights holding broadcasters who wish to use the ODF for superimpositions in their own language and other sport-specific information can produce unique TV signals for their own countries. Generation of audio descriptions from the ODF in each country's language is an effective way to use the existing ODF data and international TV signals to produce programs that everyone can enjoy.

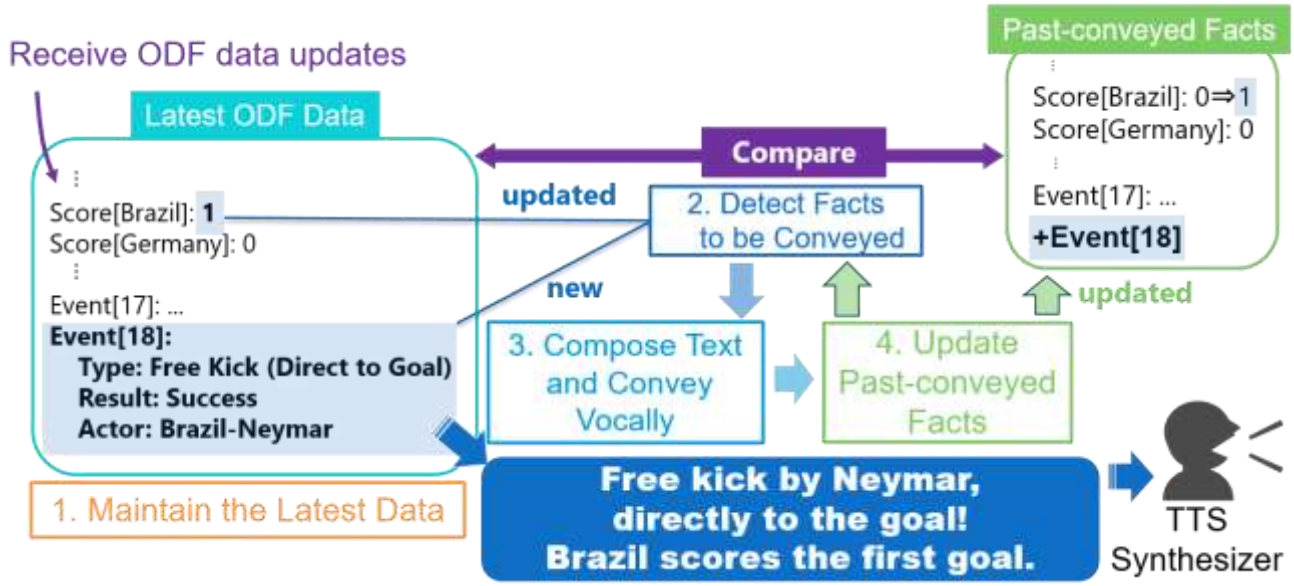

Figure 2 – Flow of Generating Commentary Text

The purpose of the generation process is to let listeners know the latest facts. In order to achieve it, we propose the following steps for generating commentary texts (Figure 2).

Step 1 - Maintaining the Latest Data

The system continuously receives updated ODF data from the venue or IBC holding the latest data. This updating process spontaneously occurs and is independent from the following process of commentary generation.

Step 2 - Detecting Facts to be Conveyed

When the system tries to compose a description worthy of conveying at that time, it first compares the latest ODF data with *past-conveyed facts*, which are maintained in the generation process. New or updated facts in the latest ODF data are the candidates to be conveyed. When no facts are detected to be conveyed, the system waits until any updates occur in the ODF data.

Step 3 - Composing Text and Conveying Vocally

The system next chooses a *template* that is best to convey the new or updated facts. Each template is a manually prepared text which has some placeholders to be filled in with a certain part of the ODF data. The complete text is then promptly conveyed vocally with a TTS synthesizer.

Step 4 - Updating Past-conveyed Facts

After the speech is conveyed, all the facts filled in the text are considered as conveyed and are kept as past-conveyed facts. Finally, the procedure goes back to step 2 to compose a new description.

Note that no new generation can be started before the previous speech conveyance has completely finished. This limitation ensures that the system generates a series of non-overlapping speech descriptions as commentary of the event. The system also has a capability to consider a specific type of past-conveyed facts as unconveyed after a certain time passes after it was conveyed. This enables the system to repeatedly convey some kinds of unchanged facts such as event information (e.g. Men's football final) and changeless scores. A set of templates and their choosing rules depending on the sort of new or updated facts have to be prepared manually for each sport event. We have developed a special programming language for easily describing them for a wide variety of events.

TEXT-TO-SPEECH SYNTHESIS

The commentary text is transformed into an audio description by means of text-to-speech (TTS) synthesis. We used the AITalk (5) system distributed by AI Inc. (a speech synthesis company in Japan) as our TTS synthesizer for the tests at the 2016 Oympic and Paralympic Games. AITalk uses a concatenative synthesis technique that selects, compiles and concatenates short samples of recorded sound. Since this method connects short speech waveforms to produce sound, the sound quality typically deteriorates when words not in the database need to be synthesized for use in arbitrary sentences.

TTS with Deep Learning

The TTS works by statistical parametric synthesis (SPS-TTS), which reconstructs speech waveforms using acoustic parameters predicted from acoustic modeling of a speech corpus. Deep learning, which is a method from the field of artificial intelligence, has recently been added to SPS-TTS (6). SPS-TTS utilizing deep learning first appeared in 2013 and is progressing at an unbelievable speed. We are planning to deliver programs with audio descriptions made using TTS with deep learning (DL-TTS) over the Internet for the PyeongChang Winter Olympic Games in South Korea in February 2018. DL-TTS has features suitable for sports commentary. This method produces more appropriate commentary speech for non-recorded words in terms of, for example, person and place names that are often the subjects of sports commentary. DL-TTS incorporates various extensibilities. SPS-TTS, however, requires a speech synthesizer (e.g. a vocoder) and suffers from such acoustic disadvantages as muffling. Furthermore, an enormous amount of appropriate speech data is needed in order to improve its sound quality.

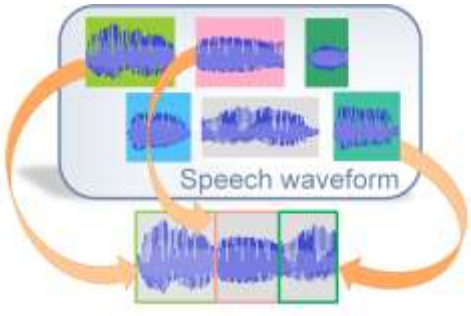

Figure 3 – Concatenative synthesis

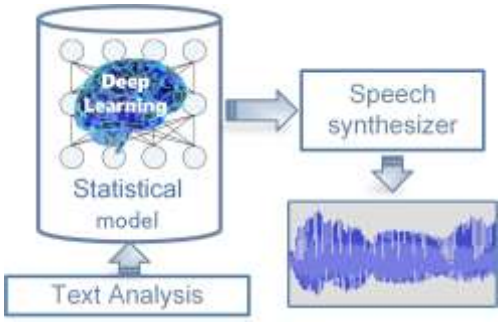

Figure 4 – SPS-TTS

TTS with Emotional Expression

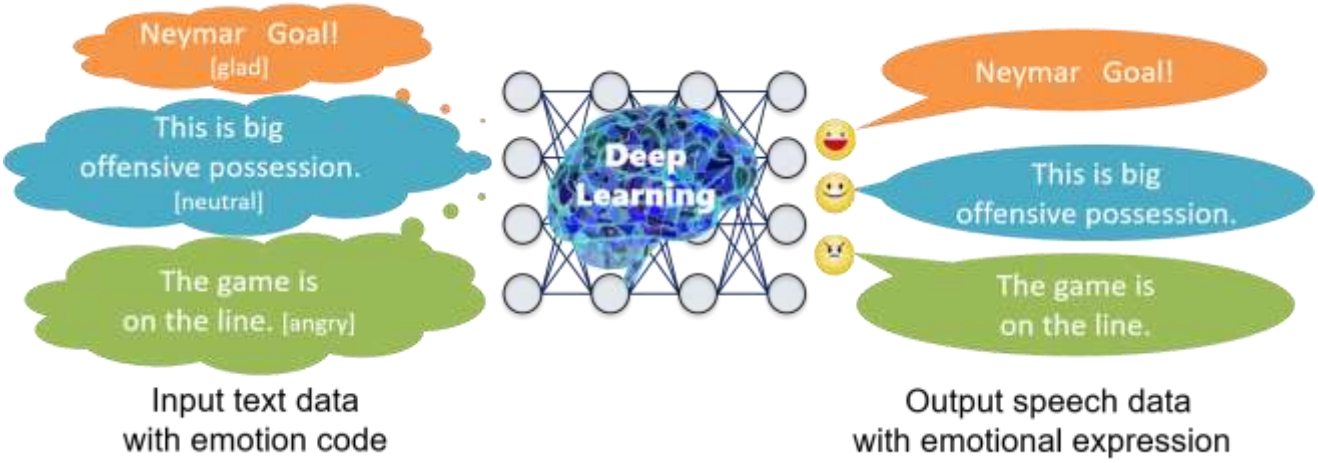

Figure 5 – Overview of DL-TTS with emotional expression

Audio descriptions with emotional expression will enhance the enjoyment of sport programs. DL–TTS with emotional expression needs machine learning to make an emotionally expressive TTS model (7). This model has a feature that outputs emotional speech when an emotion code is input. We use speech labels based on information in which emotions have been displayed for model generation. Figure 5 shows a conceptual diagram of TTS using deep learning that realizes emotional speech. It shows that an emotional code affects the final element of the model. The commentary text generation process estimates the situation of the competition from the ODF and designates the optimal emotion. It becomes possible to read out emotional speech automatically.

Latest Trends in TTS

Deep learning has been a driving force behind recent advances in TTS. Among the many new methods proposed, Google's Wavenet (8), which generates raw audio of speech data directly, is reported to sound the most natural and is now being regarded as a major breakthrough in the field of TTS. Another newcomer to the field is end-to-end TTS (9)(10), which synthesizes speech directly from a script. The framework of TTS is, therefore, expected to change. NHK is studying the various advances in these fields and intends to apply them to broadcast-related services.

EXPERIMENTS AT THE 2016 OLYMPIC/ PARALYMPIC GAMES

We conducted experiments on automatic generation of audio descriptions with offline processing at the Rio de Janeiro Olympic and Paralympic Games in 2016. In these experiments, we produced sports programs offline and used ODF and IPVandA files after delivery. Our system generated and distributed audio descriptions for 2,230 contests in 17 events at the Olympics and 266 contests in 6 events at the Paralympics (11). In so doing, we succeeded in automatically generating many programs and demonstrated the effectiveness of this experimental system.

Table 1 lists the numbers of programs produced for contests in each event.

Olympic Games			
Sport	No.	Sport	No.
Judo	414	Handball	60
Wrestling	273	Basketball	60
Boxing	255	Volleyball	60
Swimming	205	Football	56
Badminton	203	Fencing	46
Tennis	124	Canoeing	16
Table Tennis	112	Diving	12
Archery	91	Athletics	10
Beach Volleyball	86	Synchronized Swimming	5
Hockey	72		
Rugby Sevens	66	Trampoline	4
Paralympic Games			
Sport	No.	Sport	No.
Judo	74	Football	24
Table Tennis	72	Wheelchair Tennis	18
Wheelchair Basketball	67	Sitting Volleyball	12

Table 1 – Rio Olympics Programs with automatically generated audio descriptions

FUTURE PROJECTS

We plan to produce a live distribution service for the PyeongChang Winter Olympic Games, 2018. After that, we will examine the issues raised during the live broadcasts with the goal of introducing a working service at the Tokyo Olympic and Paralympic Games in 2020.

The volume of information required for audio commentary depends on whether the listener is sighted or visually impaired. Together with the development of a video production and distribution service that provides useful automatic audio commentary for sighted persons, we are proceeding with research on the audio commentaries to assist visually impaired persons in their TV viewing. Our aim as Japan's public broadcaster is to develop automatic distribution technology so that everyone, including visually impaired persons, can enjoy sports programming on TV.

In the future, a general-purpose system that supports data other than ODF will be needed for big data such as that from Twitter and Facebook and also for automatic document generation by using image recognition. Such systems will support multiple languages.

The technology for generating commentary text and TTS is advancing rapidly with the advent of artificial intelligent technologies in the form of deep learning. These technologies have the power to make audio descriptions more intelligent and more natural sounding. We will study the latest human-friendly broadcasting technologies with a firm commitment to making broadcast services more accessible to everyone.

CONCLUSION

We described an automatic audio description generation system for sports programs. This system makes it possible to attach commentary automatically to international TV signals. It has been shown to be effective not only for visually impaired persons but for sighted persons as well. Automatic generation of audio descriptions is a highly efficient way to cope with the attendant costs of sports programming.

REFERENCES

1. Ministry of Internal Affairs and Communications, 2014. Actual results of attachment of closed caption and others in 2015, http://www.soumu.go.jp/menu_news/s-news/01ryutsu09_02000126.html (in Japanese).

2. Olympic Broadcasting Services, 2017. The International Signal, About OBS, https://www.obs.tv.

3. Olympic Broadcasting Services, 2015. On-Venue Results and Timing. Broadcaster Manual Rio 2016 Olympic Games, 2015. pp. 84 to 85.

4. International Olympic Committee, 2017. Olympic Data Feed, http://odf.olympictech.org.

5. AI Inc., 2017. What is AITalk, http://www.ai-j.jp/english.

6. Zen, H., et al., 2013. Statistical Parametric Speech Synthesis Using Deep Neural Networks, In Proc. ICASSP, 2013. pp. 8012-8016.

7. Kurihara, K., et al., 2017. The DNN-based speech synthesis using speaker codes and emotional codes, IEICE General Conference 2017, 2017. vol. 2017, D-14-10 (in Japanese).

8. Van den Oord, A., et al., A generative model for raw audio, CoRR abs, 1609.03499.

9. Wang, Y., et al., 2017. Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model, arXiv preprint arXiv, 1703.10135, 2017.

10.Jose, S., et al., 2017. char2wav: end-to-end speech synthesis, ICLR 2017.

11.Miyazaki T., et al., 2017. Automatic Generation of Audio Description For Olympics / Paralympics Programs, NAB Show 2017.