Vol.2010-SLP-80 No.1
2010/2/12

音声合成を用いたインターネット動画用音声ガイド

立花隆輝[†1] 長野徹[†1]
高木啓伸[†1] 西村雅史[†1]

筆者らは、音声合成 (TTS) を用いてインターネット動画用の音声ガイド (AD) を作成・流通しやすくする枠組みを開発している。AD を TTS でどのように作成すればどれほど有用になるのかまだ明らかではなかった。AD 作成において、音声合成ならではの新しい表現方法、映画など感情表現を伴う動画での有用性、感情音声合成の効果なども興味深い疑問点である。本論文では、これらの疑問点に関して TTS を用いた AD の予備実験の結果を紹介する。そこでは、特にドラマに対して高品質な TTS の有効性が示唆された。そして、まもなく実施予定の本実験に向け準備中の感情音声合成についても実験結果を紹介する。TTS を利用することによって AD 作成の負担が軽減され、無数のボランティアが作成した多数の AD が利用可能になることが期待される。

TTS to Provide Audio Descriptions of Internet Videos

Ryuki Tachibana,[†1] Tohru Nagano,[†1]
Hironobu Takagi[†1] and Masafumi Nishimura[†1]

We are developing a collaborative Web accessibility framework that facilitates the authoring and sharing of Audio Descriptions (AD) for Internet videos by using a Text-To-Speech (TTS) engine. The crucial aspects of TTS-generated ADs and their utility are still unknown. It is natural that two-hour romantic movies would require higher quality for TTS-generated ADs than are needed for short e-learning videos. In this paper, we introduce the results of preliminary experiments of TTS-generated ADs for two video genres. The results suggest that an AD generated with a high quality TTS system is feasible for dramas. We also present experimental results for expressive TTS systems. We believe that TTS-generated ADs reduce the authoring costs and will allow for widespread sharing of ADs created by large numbers of volunteers.

1. Introduction

As the bandwidth used to access the Internet increases, online video is increasingly used by both consumers and enterprises to entertain, inform, or educate website visitors. There are various categories of online videos such as movies (including dramas and animations), e-learning materials, home videos, commercials, etc. However, in most cases, these online videos are difficult for blind or visually impaired people to understand or enjoy. This is because almost none of them include an audio description (AD), which is an additional narration track that establishes the scenes and describes the non-verbal actions so these people can understand the videos.

We are working on a project[1)] to provide a framework and tools to expedite the authoring, sharing, and use of ADs based on Text-To-Speech (TTS) technologies and collaborative Web accessibility improvements[2)]. The project puts particular emphasis on using TTS-generated ADs for various genres of videos. Before starting the project, we predicted that videos including emotional expressions, such as humorous animations or tragedies, must require higher TTS quality for the viewers to appreciate the video compared to emotionless video genres such as e-learning materials.

In this paper, we summarize the ongoing project and describe the role and the requirements of the TTS technology. We present preliminary experimental results showing that most of the subjects agree that TTS-generated ADs based on a state-of-the-art TTS system are usable even for dramas with emotional expressions. We also present results of other experiments with expressive TTS systems, which we are preparing for the final experiments of the project.

2. Project Background, Challenges and Questions

Our surveys on the availability of ADs and captions for online videos discovered that ADs are far less available than are captions. The reasons for this seem to include (1) it requires special skill to write AD scripts that adequately and com-

†1 日本ＩＢＭ東京基礎研究所
IBM Research - Tokyo

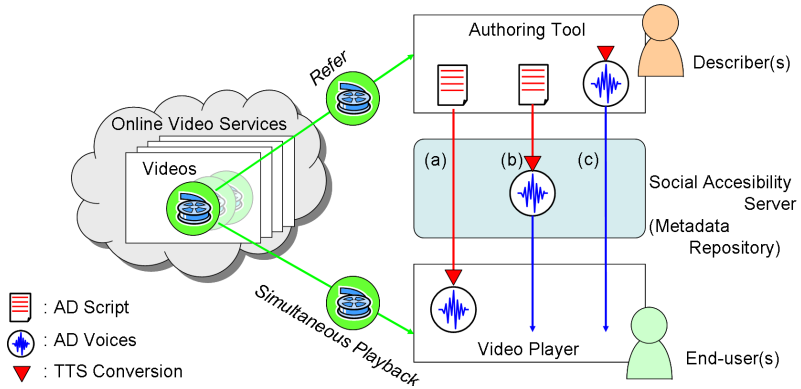

Fig. 1 The Project Framework.

pactly describe the scenes, and (2) recording ADs demands professional narrators or recording by the same person who is creating the description.

Sharing

To overcome these difficulties, our project is developing a framework based on the idea of collaborative Web accessibility improvements[2]. In the framework, non-expert volunteer describers write the AD scripts for the online videos, and the TTS engine converts the scripts into ADs. The end-users play a video and the corresponding TTS-generated AD simultaneously by using a dedicated player. There are three options regarding where the TTS conversion is done: (a) in real time on the PCs of the end-users (listeners), (b) at an intermediate server, and (c) on the PCs of the describers. The advantages of (a), the live TTS approach, include that the end-users can use the TTS system they are familiar with and that they can freely control the synthesis when playing the AD-equipped Videos (ADVs). In contrast, both (b) and (c), the stored-audio approaches, let end-users without a TTS system on their local PCs or with more limited playback device enjoy the ADVs. The scenario of (b) allows use of a powerful server-based high quality TTS system. In the scenario of (c), the describers can minutely adjust the time alignments of ADs. Each of the scenarios has advantages, and we are planning to support all three scenarios in our framework. It should be possible to share the ADs in the user's preferred form whether as an AD text script or as an AD voice. The server and the file format for ADs should support both.

Authoring

Writing AD scripts is not an easy task. There are often many actions and objects on the screen that must be described. However, if ADs collide with the characters' speech or other meaningful sound events in the original video, then understanding the video is even more difficult. Hence, it is usually quite challenging for the describers to create compact descriptions and find gaps into which ADs can be inserted. To make the AD authoring easier, our project is developing an authoring tool in which the time line is shown so that the describer can find gaps for the ADs. Since the tool is connected with a TTS system through its API, the describer can easily produce synthetic voices by inputting the AD scripts into the authoring tool.

Playback

At playback time, precise synchronization between the ADs and the video is required to avoid collisions. This is a challenge for the live TTS approach since it must play the freshly synthesized AD without any delays that could cause collisions. An option to avoid the collisions, other than simply reducing ADs, is to stop the playback of the video while the ADs are being played, especially when there is not enough room for AD playback. A variation of this option, which may be effective for interactive video players, is to prepare two classes of ADs, default ADs and supplemental ADs. By default the player would hear the AD but not the supplemental AD, but there would be a distinctive audio signal when a supplemental AD is available. If the end-user pushes the appropriate button, then the player stops and plays the supplemental AD. However, the tempo of a video is sometimes an important aspect of appreciating that video. Another possible option is to speed up the ADs. Many blind people using screen readers often prefer screen readers that speak quite rapidly, so it may be feasible to use accelerated synthetic voices for the ADs. Flexible duration (speed) adjustment of the synthetic voices is important for this purpose and also for avoiding collisions.

Video Genres

The desirable characteristics of TTS-generated ADs may differ depending on the genre of the video. It is natural that robotic synthetic voices would be acceptable for short e-learning videos, but not for two-hour romantic movies.

More humanlike higher quality that does not interfere with the viewing must be required for some genres. It is true that many blind people are accustomed to listening to synthetic voices from screen readers, but a major difference between screen readers and ADs is that, while the user's ears are focused solely on the synthetic voice when using a screen reader, the synthetic voice and the characters' voices are alternating when listening to an ADV. This means the synthetic voice is constantly contrasted with that the human voices. All of this makes ADs a challenging task for a TTS system.

Speaking Style

When human narrators record ADs, they change the tones of their voices depending on the scenes in the drama. For example, they read the ADs with a sad feeling for sad scenes. Though there is disagreement about how much expression a narrator should add to an AD, slight changes of tone tend to occur naturally even if the narrators is avoiding dramatic expressions. Expressive TTS is a technology that can produce synthetic voices with emotions, such as happiness or sadness. Changing the emotions of synthetic voices depending on the scenes may improve the viewing experience of the end-users.

Questions and Challenges

In summary, here are some of the questions and challenges regarding TTS for ADs that will be addressed in this project:

- Are TTS-generated ADs acceptable?
- What aspects of a TTS system are important for ADs?
- Is expressive TTS useful for ADs?
- How should a TTS-generated AD be played?
- What are the best methods for writing AD scripts for TTS?
- How do the genres of videos affect TTS-generated ADs?

3. Our TTS Systems for Audio Description

In this section, we describe the TTS systems we are using in the experiments for this project. Note that the project does not require the use of these specific systems. The framework and the tools developed within the scope of the project are also designed to work with other TTS systems. Since we are focusing on the use of high quality TTS systems to avoid spoiling the viewing experiences of emotional videos, we chose the stored-audio approach using a high-quality server-based TTS system.

3.1 Basic TTS System

Our basic system for the experiments is a unit-selection TTS system[3)–5)]. The minimum unit of concatenation is a sub-phoneme. The units are clustered by a context-dependent decision tree whose window size is five phonemes. Though it is possible to reduce the number of units by preselection, we did not do preselection in this project. The run-time process searches for the unit sequence with the minimum cost, performs PSOLA to adjust the unit durations and the F0 gaps, and concatenates them. The output of the text processing module was manually corrected for the experiments. Our basic corpus for the project consists of approximately 6,900 sentences, which is approximately 7.6 hours of speech (after silence removal). We collected the corpus by asking a narrator to cheerfully read sentences about car navigation, weather, news, etc.

3.1.1 Expressive TTS System

This unit-selection system is also capable of producing emotional synthetic voices. It requires a middle-sized corpus for each of the emotions to be reproduced as well as the large basic corpus. We used two emotional corpora in Japanese, one for conveying good news, one for bad news, and one for excitement. The good-news and bad-news corpora each consist of approximately 1,000 sentences, which is approximately 1.2 to 1.3 hours of speech. An emotional corpus was collected by asking the narrator to repeat all of the scripts with the target emotion. An examples of a good-news sentence might be "You have won special pajamas and a T-shirt with the exclusive logo of this TV program."

The speech units generated from the emotive corpora are labeled with the corresponding emotion. We also built prosody models for each of the emotions. When generating synthetic voices with an emotion, the speech units whose emotive labels are different from the target expression also become candidates for concatenation, but with predefined penalties. In this way, only when the search cannot find a speech unit with the target emotive label and prosody values, then a speech unit with a different emotive label is used as an alternative.

3.2 Conventional TTS System

For comparison and also for rapid testing of the authoring tool, we also used

Voice	G1	Human narrator	G2	High quality TTS	G3	Conventional TTS
Playback	P1	Normal amount Normal speed	P2	Double size Double speed	P3	Double size Interrupt
Genre	C1	E-learning	C2	Drama		

Table 1 Parameters for the preliminary TTS-generated AD experiments.

a conventional Japanese TTS system[6], which is widely used with the popular screen readers[7] and[8]. This system uses little memory and has high performance, though the voice quality is not as good as the state-of-the-art TTS technologies. The system is a concatenative system with a limited number of speech units, each of which is used for a specific phonetic context by performing signal processing.

4. Experiments

To answer the questions listed in Section 2, we are conducting a series of experiments.

4.1 Preliminary TTS-generated AD Experiments

First, to obtain rough understanding of the problem space, we performed an experiment in which we gathered detailed comments from a small number of subjects regarding TTS-generated ADVs. For the experiments, we used three methods to generate the ADs: (G1) a female human narrator, (G2) the high quality TTS system (Sect. 3.1), and (G3) the conventional TTS system (Sect. 3.2). Three playback methods were used: (P1) playing simple descriptions at a normal speed, (P2) playing approximately double-size descriptions at double speed, and (P3) playing approximately double-size descriptions at normal speed by stopping the playback of the video when the gap is too short. We also sped up G1 by changing the durations of the voices with Sony SoundForge. The genres of the videos were (C1) e-learning (cooking instructions) and (C2) drama (excerpts from a movie). By combining these configurations, we generated 18 test samples (Table 1), each of which was approximately two minutes long, and played all of them for each of the subjects. The original video without an AD was played before the playback of each test sample for comparison. The test subjects were three blind or visually impaired people and a sighted person. For each of the samples, we asked the subjects to give comments and to rate it using some features, though the number of subjects was too small for statistical analysis. The most important feature

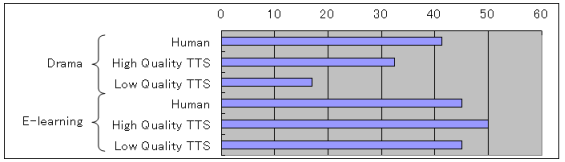

Fig. 2 The goal improvements.

was goal achievement. The goals for C1 and C2 were set to be able to cook the dish and to enjoy the story, respectively.

Results

Since the subjects were used to listening to screen readers, there was almost no intelligibility problem. Some reported the poor quality of G3 caused a slight surprise at the beginning of the test sample and that it made them tired. However, all of them said that the voice quality did not matter too much for C1 and that the AD's availability is more important. For G2, the dominant opinion was that its more natural intonation and voice quality made viewing easier. The effect of the high quality of G2 for C2 can be seen also in Figure 2, which shows the average values of the differences between the goal achievement scores for playback with the AD and without the AD. However, there were a few dissatisfied comments about the incomplete smoothnessof G2.

Regarding the playback methods, though the subjects reported that they could understand the double-speed ADs, in general they did not like P2 since it required too much concentration. For C2, P1 was chosen as the best because its preservation of the tempo of the story was more important and because less information was required to understand the story. Some reported the abrupt pause of the background sounds caused by P3 was annoying. In contrast, for C1, P3 was the favorite. P1 apparently failed to supply the necessary information for the recipe.

The subjects also mentioned the importance of precise synchronization of the ADs and the necessity for a proper volume balance between the original sounds and the ADs.

4.2 Subjective Listening Tests of Expressive Japanese TTS

To ensure that the system is capable of producing different speaking styles and that the acoustic qualities for the emotional styles are useful, we conducted

subjective listening tests. A total of 11 subjects participated in these tests. To learn the speaking styles, the subjects were first exposed to five samples from both the neutral and the emotional corpora. Then, each of the subjects was asked to rate test samples in eight categories. Three categories were again the human voices from (HN) the neutral corpus, (HG) the good-news corpus, and (HB) bad-news corpus. One category (TBN) was for the neutral TTS system (Sect. 3.1). Two categories were (TEG) good-news samples and (TEB) bad-news samples generated by the emotional TTS system (Sect. 3.2). For comparison, we produced a basic TTS system based on each of the middle-sized emotional corpora (TBG for good-news and TBB for bad-news). We synthesized over 80 sentences with versions for each of the TTS systems. For human voices, we chose relatively emotion-free sentences from the corpora to make it difficult for the subjects to guess the categories based on the meaning of the sentences. A test set was generated for each of the subjects by randomly choosing five test samples for each of the eight categories, which resulted in 40 test samples in total for each of the subjects. The order of the categories in each set was random and we did not inform the subjects which voice was from which category. For each test sample, we required the subjects to rate the acoustic quality by selecting answers from 5 (Very Natural) to 1 (Very Unnatural) and also to classify them into emotions (Neutral, Good-news, Bad-news, and Other).

Results

Table 2 shows the experimental results. Figure 3 is a graph of the MOSs (Mean Opinion Scores) for the acoustic qualities. First, by comparing TEG, TBG, and TBN in Table 3 , we can see that the expressive TTS system (TEG) produced a far higher score than the basic good-news system (TBG) did and that the quality of TEG was as good as that of the basic neutral system (TBN). However, the good-news voices produced by TEG were often perceived as neutral (36.4%) and the neural voices produced by TBN were often perceived as good-news (49.1%). Furthermore, the subjects sometimes confused HN with HG. This is because the neutral corpus was itself composed of cheerful readings and it was difficult for the narrator to produce more cheerful tones. Since the neutral units and the good-news units were not very different, they were interchangeable.

In contrast, for the bad-news style, the basic bad-news system (TBB) was much

	TTS System	Target Category	MOS	Perceived Category (%)			
				Good	Neutral	Bad	Other
HN	Human	Neutral	4.46	23.6	**72.7**	1.8	1.8
HG	Human	Good	4.56	**85.5**	12.7	1.8	0.0
HB	Human	Bad	4.60	0.0	36.4	**63.6**	0.0
TBN	Basic	Neutral	3.13	49.1	**49.1**	1.8	0.0
TBG	Basic	Good	2.60	**41.8**	52.7	1.8	3.6
TBB	Basic	Bad	2.95	1.8	25.5	**72.7**	0.0
TEG	Expressive	Good	3.18	**61.8**	36.4	0.0	1.8
TEB	Expressive	Bad	2.91	9.1	38.2	**41.8**	10.9

Table 2 Results of subjective listening tests.

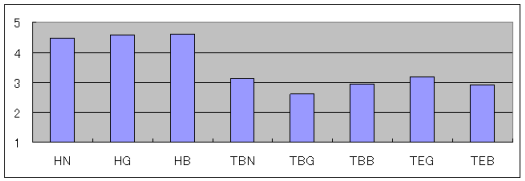

Fig. 3 The Mean Opinion Scores for the acoustic quality.

better than the expressive system (TEB). Since the prosody values predicted by the bad-news prosody models were different from those of the neutral units, the neutral units were not used very often for TEB and mixing neutral units among bad-news units sometimes resulted in a perception of inconsistent speaking styles within one test sample. We can also see that, with low F0 values and a narrow F0 range, the bad-news style (TBB) was easier to synthesize compared to the good-news style (TBG) when the corpus size was middle. Based on these results, it can be said different approaches would be best for the good-news style and the bad-news style because of the characteristics of the training corpora.

5. Related Works

There has been little research on TTS-generated ADs. Sato's group developed a mouse-like playback interface for TTS-generated ADs[9]. Their main interest was in interaction design and they subjectively compared several different playback methods. Gagnon presented software tools for authoring and playing of TTS-generated ADs[10]. This authoring tool uses various video processing technologies such as scene detection and text recognition to reduce the authoring costs. The

playback tool is an interactive tool using a live approach and the end-users can use the tool to read the necessary information at any time while watching the video. Though these projects share similar motivations and targets with our work, the voice quality of the TTS and video genres were not discussed in these projects.

For emotional TTS, a style adaptation approach[11),12)] that converts a neutral speaking style to an emotional style is a more standard approach and has the advantage of requiring a smaller emotional corpus than our approach does. The fast performance and compactness of HMM-based synthesis systems may fit very well with the live approach of TTS-generated AD playback. However, in our understanding, unit-selection systems still have advantages for the acoustic quality for in-domain sentences, and especially when a large database can be used at the server. We wanted to cover various TTS engines in our framework.

6. Conclusions

In this paper, we briefly described a framework to facilitate the authoring and sharing of ADs by using TTS. We clarified some TTS requirements by reviewing the sharing, authoring, and playback procedures of TTS-generated ADs. There are still many open questions. To answer some of these questions and to obtain a better understanding of the problem space, we conducted preliminary experiments with TTS-generated ADs. The experiments suggested that TTS-generated ADs were acceptable, that double-speed playback was not very useful for ADs, and that the TTS voice quality was critical for ADs used for dramas while having little effect on e-learning videos. The usability of expressive TTS for ADs is also an open question. We performed subjective listening tests for the expressive TTS systems. The test results clarified the effects of the speaking styles of our emotional corpora. Since our "neutral" corpus was itself recorded with a cheerful speaking style, different approaches seem to be reasonable for the good-news style and the bad-news style. Now we are planning to conduct another set of experiments involving TTS-generated emotional ADs. We hope that collaborative Web accessibility improvement will be widely used and contribute to changing the access environments of users with disabilities worldwide.

Acknowledgments This project is being funded partly by the National Institute of Information and Communications Technology (NICT), Japan. We thank WGBH for their help in the user studies. We also thank the other members of our project.

References

1) M.Kobayashi, K.Fukuda, H.Takagi, and C.Asakawa, "Providing synthesized audio description for online videos," in *Proc. ASSETS 2009*, Oct. 2009, pp. 249–250.
2) H.Takagi, S.Kawanaka, M.Kobayashi, D.Sato, and C.Asakawa, "Collaborative web accessibility improvement: Challenges and possibilities," in *Proc. ASSETS 2009*, Oct. 2009, pp. 195–202.
3) J. F. Pitrelli, E. M. Eide R. Bakis, R. Fernandez, W. Hamza, and M. A. Picheny, "The IBM expressive text-to-speech synthesis system for American English," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1099–1108, Jul. 2006.
4) T.Nagano, S.Mori, and M.Nishimura, "An N-gram-based Approach to Phoneme and Accent Estimation for TTS," *IPSJ Journal*, vol. 47, no. 6, pp. 1973–1981, 2006, (in Japanese).
5) R. Tachibana, T. Nagano, and M. Nishimura, "F0 Gradient Model for Acoustic Quality and F0 Consistency of Concatenative TTS," in *Technical Report of IEICE, 2007-NLC/SP-12*, Dec. 2007, pp. 253–258.
6) T.Saito, M.Sakamoto, Y.Hashimoto, M.Kobayashi, M.Nishimura, and K.Suzuki, "ProTALKER: a Japanese Text-To-Speech System for Personal Computers," Tech. Rep. RT0110, IBM TRL Research Report, Jun. 1995.
7) "IBM Home Page Reader," 2005, http://www-06.ibm.com/jp/accessibility/solution_offerings/hpr/index.html, in Japanese.
8) "IBM Easy Web Browsing," 2004, http://www-06.ibm.com/jp/accessibility/solution_offerings/EasyWebBrowsing.html, in Japanese.
9) D.Sato, H.Takagi, and C.Asakawa, "An experimental interface to present audio description of video for the blind," in *Proc. WISS 2006*, Dec. 2006, pp. 71–76, (in Japanese).
10) L.Gagnon, "Automatic detection of visual elements in films and description with a synthetic voice - application to video description," in *Proc. Vision 2008*, July 2008.
11) T. Nose, J. Yamagishi, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 9, pp. 1406–1413, Sep. 2007.
12) J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Speech, Audio & Language Process.*, vol. 17, no. 1, pp. 66–83, Jan. 2009.