INTERSPEECH 2017: Show & Tell Contribution
August 20–24, 2017, Stockholm, Sweden

WebSubDub – Experimental system for creating high-quality alternative audio track for TV broadcasting

Martin Grůber[1], *Jindřich Matoušek*[2], *Zdeněk Hanzlíček*[1], *Jakub Vít*[2], *Daniel Tihelka*[1]

[1]New Technologies for the Information Society (NTIS), [2]Department of Cybernetics
Faculty of Applied Sciences, University of West Bohemia, Czech Republic
{gruber,zhanzlic,dtihelka}@ntis.zcu.cz, {jmatouse,jvit}@kky.zcu.cz

Abstract

This paper deals with a presentation of an experimental system (called *WebSubDub*) for creating a high-quality alternative audio track for TV broadcasting. The system is used to create subtitles for TV shows in such a format which allows to automatically generate an alternative audio track with multiple voices employing a specially adapted TTS system. This alternative audio track is intended for televiewers with slight hearing impairments, i.e. for a group of televiewers who encounter issues when perceiving the original audio track – especially dialogues with background music, background noise or emotional speech. The system was developed in cooperation with Czech television, the public service broadcaster in the Czech Republic.
Index Terms: creating subtitles, alternative audio track, speech synthesis

1. Introduction

Speech synthesis technologies are nowadays used in various areas of normal human life, e.g for reading e-books, as screen-readers, in car navigations or in cell phones, i.e. as a part of natural interaction between humans and devices. These technologies can however be used also for helping humans with various disabilities like those with voice/speech disorders [1] or visual [2] or hearing impairment. It was found that some televiewers do not understand dialogues in TV shows correctly because they are too fast, too lively, too emotional or there is a high level of background noise or background music. For example, senior televiewers or televiewers with hearing impairment cannot even watch some shows because of these perceiving troubles.

This paper presents a newly developed web-based application *WebSubDub* which is intended to be used for creating subtitles for TV shows and for generating alternative audio tracks using our TTS system ARTIC [3]. The alternative audio track contains only the transcribed dialogues from the show pronounced by calm voices in contrast with the original audio track which may contain background noise or emotive utterances. The alternative audio track can be more suitable for televiewers suffering from the aforementioned perceiving troubles.

Compared to other applications for creating subtitles, e.g. FAB Subtitler (www.fab-online.com), *WebSubDub* allows to manage characters in the show and to assign various utterances in the subtitles with various characters. These characters can be then assigned with various voices available on a TTS server (either manually or automatically) and thus the alternative audio track may contain multiple voices. It is obvious that the number of available TTS voices may (and usually will) be lower than the number of characters appearing in the show. Thus the same TTS voice may be used for various characters within a particular show.

The application further allows to immediately play an audio sample corresponding to each utterance and/or subtitle to check if it is generated correctly as the text may contain various non-standard expressions like numbers or abbreviations. Thus, any transcription error can be immediately fixed in the text or a rule may be created to pronounce a particular expression in a specific way. The user is also instantly informed if the length of the generated audio sample does not fit the time slot defined for the particular subtitle so that the text of the related utterance can be adjusted. Otherwise the generated speech might be sped up too much during further processing to fit the time slot and therefore the quality might be deteriorated. The whole final alternative audio track (which may contain additional optimizations, see Section 2.5) can be also generated from the application and it is then available for download or for listening to in sync with the video track.

The system architecture is described in Section 2 and the application frontend (GUI) in Section 3. In Section 4, the employed TTS system is briefly presented.

2. System architecture

The developed web application is based on a client-server architecture, running on *nginx* HTTP server, *Gunicorn* application server and with *MongoDB* database system. For speech synthesis, a separate TTS server is used.

Such architecture has several advantages including:

- accessibility from any place where a fast and stable internet connection is available;
- requirement for only a common modern web browser, i.e. the application runs on any device like PC or tablet; installation of any other software is not necessary (the only requirement is thus a possibility to install a web browser and sufficient performance to play audio and video streams from the internet);
- platform independence, i.e. the application runs independently on the device operating system;
- availability of the same most up-to-date version of the application to all users all the time;
- a centralized storage for the multimedia content (video tracks, alternative audio tracks), subtitles and other content (e.g. pronunciation dictionaries).

The application server serves mainly as an API for the client-side GUI. However, it provides also additional functionalities which are briefly presented in the following subsections.

2.1. Video conversion

The TV shows may be distributed in various video formats and encoded with various codecs. When a TV show is inserted into the system, the video track is converted into at least 2 different video formats: *mp4* (*H.264* video codec and *AAC* audio codec)

and *webm* (*VP8* video codec and *Vorbis* audio codec). Thus, all modern web browsers are able to play the video tracks.

2.2. Subtitle conversion

As the application allows importing subtitles from external sources (to be able to just modify existing subtitles), algorithms for converting subtitles between the following subtitles formats were implemented (not all conversions are lossless):

- EBU STL (developed by the European Broadcasting Union);
- SubRip (SRT);
- MicroDVD (SUB);
- ESUB-XF (European Subtitle Exchange Format developed by FAB which is internally used by the application as it is XML-based and thus very flexible).

2.3. Cut detection

To be able to do smart automatic assignment of voices to characters (see Section 2.4) and to automatically optimize the subtitle time slots (see Section 2.5), detection of cuts in the video is necessary. See [4] for more details.

2.4. Smart automatic assignment of voices

As it was mentioned before, the number of available TTS voices does not need to be sufficient to differentiate all characters appearing in a show. Thus, some of the characters must share the same voice. In the application, the user can either select a voice for each character in the particular show, or the system can automatically assign all the characters with available voices in such a way that the chance of the same voice to be used in a single scene by different characters is minimized. The detected cuts and the subtitle timing is used for that.

2.5. Subtitle time slot optimizations

The users are supposed to align the subtitle time slots according to the real speech (i.e. to the original audio track). However, the real speech can be fast whereas the TTS produces speech at a standard rate. Thus, in some situations, the time slot for a particular subtitle may be too tight, and the generated audio might not fit the time slot. Then, the audio is sped up to match the slot. In order to reduce the quality deterioration, the system tries to shift the subtitle timing in order to increase the time slot size for the generated audio. During the shifting, it is ensured that subtitles do not overlap and do not cross scenes' boundaries (which are determined on the basis of the detected cuts).

3. Frontend

The frontend (screenshot is depicted in Figure 1) is an interface where users can open and play TV shows and create subtitles. It allows the users to easily adjust the timing and to see immediately if the synthesized utterance fits the time slot or not so that they can adjust the subtitle text. They can also play the synthesized utterance and thus immediately hear whether the text is synthesized properly or not. In case there is any mispronunciation (e.g. because of expressions in foreign languages), the users may define the correct pronunciation either directly in the subtitle text (to apply the rule locally) or using pronunciation dictionaries (to apply the rule globally for all subtitles in the particular show).

The final alternative audio track can be generated directly in the frontend GUI and can be downloaded or played in sync with the original video track.

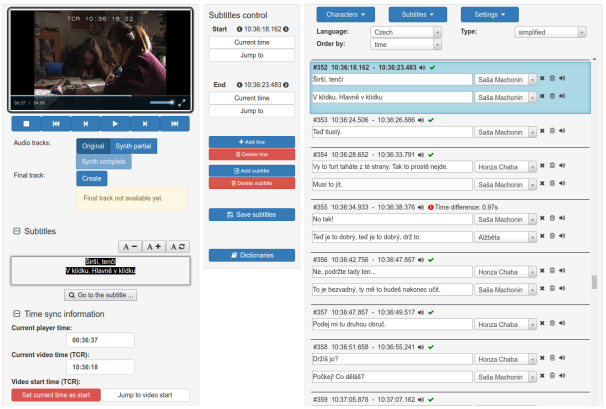

Figure 1: *Screenshot of the frontend.*

4. Text-to-speech

For synthesizing the subtitles texts and also the final alternative audio track, our TTS system ARTIC [3] utilizing both unit selection and statistical parametric methods is employed and available as a web service. It manages several Czech male and female voices as well as some voices for other languages.

The text to be synthesized is a subject of a thorough analysis and preprocessing. The local pronunciation exceptions as well as global pronunciation dictionaries created by the users are taken into consideration during the text preprocessing.

5. Conclusions and future work

This paper presents an application which serves for creating TTS-friendly subtitles and employs TTS technologies for automatic generation of an alternative audio track for TV shows from the subtitles. Czech television plans to broadcast the alternative audio track together with the original track in the future.

Our future work will be focused on generating an audio track containing description of scenes (*audio description*) which is used by visually impaired televiewers.

6. Acknowledgements

This research was supported by the Ministry of Education, Youth and Sports of the Czech Republic, project No. LO1506 and by the grant of the University of West Bohemia, project No. SGS-2016-039.

7. References

[1] M. Jůzová, J. Romportl, and D. Tihelka, "Speech corpus preparation for voice banking of laryngectomised patients," in *Text, Speech, and Dialogue*, ser. Lecture Notes in Computer Science. Springer, 2015, vol. 9302, pp. 282–290.

[2] M. Grůber, J. Matoušek, Z. Hanzlíček, Z. Krňoul, and Z. Zajíc, "ARET — automatic reading of educational texts for visually impaired students," in *Proceedings of Interspeech*, San Francisco, California, USA, 2016, pp. 383–384.

[3] J. Matoušek, D. Tihelka, and J. Romportl, "Current state of Czech text-to-speech system ARTIC," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, vol. 4188, Springer, 2006, pp. 439–446.

[4] J. Matoušek and J. Vít, "Improving automatic dubbing with subtitle timing optimisation using video cut detection," in *Proceedings of ICASSP*, Kyoto, Japan, 2012, pp. 2385–2388.