Towards Automatic Audio Track Generation for Czech TV Broadcasting: Initial Experiments with Subtitles-to-Speech Synthesis

Zdeněk Hanzlíček, Jindřich Matoušek and Daniel Tihelka
University of West Bohemia, Faculty of Applied Sciences, Department of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
{zhanzlic, jmatouse, dtihelka}@kky.zcu.cz

Abstract

In this paper, the project "Elimination of the Language Barriers Faced by the Handicapped Watchers of the Czech Television" aimed at making Czech TV broadcasting available to a broader group of TV watchers is introduced. More specifically, the problems of the automatic audio track generation within the project are mentioned. As the audio track will be produced from subtitles, text-to-speech (TTS) technology will be utilised. Several versions of a TTS system planned to produce the audio track are described. In this paper, the main attention is paid to the analysis of synchronicity between subtitles and the synthetic speech. Problems with fitting synthetic speech into the predefined subtitles slots were revealed – for more than 44 % of all subtitles, the synthetic speech overlapped the slots. So, great care will have to be taken to produce speech at faster rates when customising our TTS system for the task of generating audio tracks from subtitles.

1. Introduction

This paper presents the first steps towards the automatic generation of audio tracks from subtitles in the ELJABR project. ELJABR is a Czech acronym for "Elimination of the Language Barriers Faced by the Handicapped Watchers of the Czech Television". The aim of the project is to make Czech TV broadcasting available to broader group of TV watchers. The project is solved in cooperation with Czech Television. Within the project, two main tasks are researched. The first one is automatic real-time subtitling of speech in live TV broadcasting [1]. It is aimed especially at the deaf or hearing impaired TV watchers. The second task is the automatic generation of the audio track from existing subtitles. This service is planned to be used by watchers with minor hearing impairments like seniors, people with dyslexia or minor mental retardation.

This paper concerns the latter task. As most of the TV programmes are provided with subtitles (or closed captions, see Section 3), this information (mostly broadcasted as a plain text using a teletext page, typically 888) is used as an input to a text-to-speech (TTS) system customised to this task and a new audio track ("audio subtitles") is produced in a fully automatic way. As a result, a TV programme could be supplemented with another audio track. The track is less dynamic, more undisturbed and is supposed to be helpful for the aforementioned groups of TV watchers, and also for people who simply are not able to follow the complex sound structure of modern TV programmes - they do mind the lower intelligibility of real dialogues or possibly also music or effect component present in the original audio track. Every TV watcher will then be able to choose between the original and the TTS-generated audio track.

The paper is organised as follows. In Section 2, several versions of a TTS system planned to generate the audio tracks are described. In Section 3, subtitles issues and the experimental subtitles available are specified. Section 4 describes the first experiments with synthesising speech from subtitles and analyses the problems of fitting the TTS-generated speech to the predefined subtitles slots. In Sections 5 and 6, the results and solutions to the problems are discussed, and future work is outlined.

2. TTS system ARTIC

A Czech TTS system ARTIC (Artificial Talker in Czech) is used to generate the accompanying audio track from subtitles in a fully automatic way. Two versions of speech synthesis system are currently supported: single unit instance (SUI) system and multiple unit instance (MUI) system [2]. Both versions

utilise corpus-based concatenative speech synthesis technology. The single unit instance system uses a compact acoustic unit inventory (there is only one instance of each speech unit present in the inventory) and thus it is more suitable for low-resource devices (mobile phones, pocket PCs, etc). The SUI system is more flexible from the point of view of signal modifications - in order to meet the explicit prosodic specifications (F0 contour, duration and intensity) estimated from the input text, each speech unit is a subject of a signal modification. In this approach, duration of each speech unit could be easily modified (and especially shortened) so that synthetic speech of the generated audio track does fit into the time slots predefined in the source subtitles.

On the other hand, the multiple unit instance system takes more instances of each speech unit into account and selects the optimal instances dynamically during synthesis runtime (using a unit selection technique) [4]. Consequently, the resulting synthetic speech is of a higher quality, but at the expense of enormous memory requirements. The MUI version of our system utilises an implicit prosodic specifications - the required prosodic characteristics are described only at a high-level by symbolic features (such as the type of prosodeme the unit is in, the order of the unit in a word, the phonetic context of the unit, etc.) - and the unit instances are selected accordingly. As it is generally known that any signal modification can cause some degradation of the output speech, our current implementation does not allow any modifications of the output speech signal. Thus, duration of TTS-generated speech is currently out of control in this approach.

Since the highest possible quality of the synthetic speech is required in the ELJABR project, the MUI version is planned to be used and adapted for the needs of the project. In this paper, however, the SUI system is also used to analyse the issues of fitting the TTS-generated speech into the predefined time slots in the subtitles.

As for the synthetic voices, two brand new voices (one male and one female) are planned to be built within the ELJABR project. The male voice has been already built following the methodology described in [5]. The TTS system running a MUI version with this voice will be referred as MM henceforth. For the purposes of measuring the fitting the TTS-generated speech into the subtitles time slots other "older" voices were used as well, resulting in another three versions of the TTS system: a female voice and MUI engine (MF), the same female voice and SUI engine (SF), and a male voice (other than the one for MUI) and SUI engine (SM).

3. Description of subtitles

In the TV broadcasting, subtitles (also known as closed captions, or subtitles for the hearing impaired) could be viewed as an extra service especially for the hearing impaired which supplements the standard video and audio tracks with a transcript (although not always verbatim) of the audio track. The subtitles present the only source of information which could be exploited when generating the accompanying audio track for TV broadcasting. Czech Television currently broadcasts the subtitles using a teletext page 888.

At present, the EBU Subtitling Data Exchange Format [6] is used for storing subtitles of particular programmes in the binary data files. Each file (which usually ends with extension ".STL") comprises one General Subtitle Information (GSI) block followed by a number of Text and Timing Information (TTI) blocks. In the GSI block, information on the overall program is defined, such as original and translated title of programme and episode, original language, author and some rather technical information for broadcasting and display. Each TTI block defines one subtitle – it is given by a subtitle text and its start time ("in-cue"), finish time ("out-cue"), position on screen, etc.

For initial experiments, a set of 20 subtitle files (5794 subtitles in sum) for various programmes was available. It comprised several documentaries, talk-shows, fairy-tales, cartoons and miscellaneous movies.

4. Experiments

Each subtitle is defined by a text and a time slot. By the analysis of available subtitles we have found out that the dependence between the subtitle text length (in phonemes) and its time slot duration is quite weak; it is depicted on Fig. 1. The correlation coefficient is 0.37.

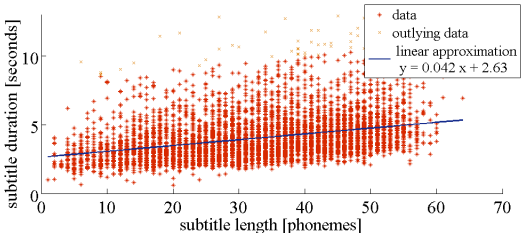

Fig. 1. Relation between subtitle text length (in phonemes) and its duration.

Usually, the corresponding synthesised utterance generated by a TTS system does not exactly fit into the given subtitle time slot. That could conduce to complications during audio track generation and

potential desynchronisation between the programme and synthesised audio track.

No serious problem arises in case the utterance duration is shorter or equal to the subtitle time slot length or also in case the utterance exceeds the slot but it does not overlap into the following subtitle time slot (see Fig. 2). The correspondency between subtitle and utterance starts seems to be crucial for an easygoing programme watching and a simple orientation in dialogues, whereas a discrepancy between subtitle and utterance ends is usually not so important.

On the other hand, a problem arises when a synthesised utterance exceeds the length of given subtitle time slot and overlaps into the following slot. Then the following utterance must be delayed and a significant audio track desynchronisation occurs. In some cases (in fast dialogues) the delay accumulates and the desynchronisation deteriorates (see Fig. 3).

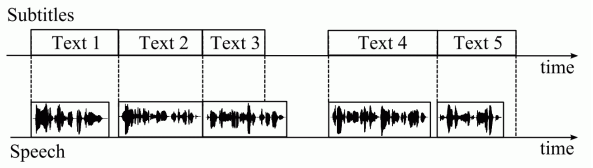

Fig. 2. Ideal conditions for synthesis – no synthesized speech desynchronisation occurs.

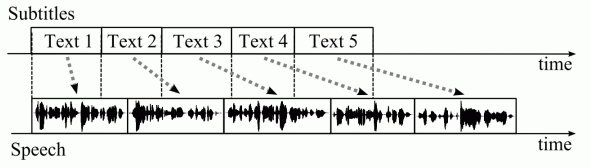

Fig. 3. The origin of the synthesized speech desynchronisation.

In our experiments, all version of TTS systems (see Sect. 2) were employed to synthesise all the subtitles. We assessed how often the desynchronisation between the synthesized audio track and subtitles occurs. Since the total number of subtitles was limited, all statistics were calculated for all programme genres together.

Tab. 1. Comparison of subtitle duration (SubDur) and corresponding synthesized utterance duration (UttDur).

	VM	MM	MF	SM	SF
SubDur > UttDur [%]	61.2	53.3	51.9	77.1	86.3
SubDur < UttDur [%]	38.8	44.7	48.1	22.9	13.7
Avg. sub. overlap [sec]	0.78	0.95	1.01	0.58	0.44

In Table 1, the difference between the given subtitle time slot length and the duration of corresponding utterance is analysed. Obviously, the result depends on used voice and version of TTS system. Detailed results for the MM synthesiser (which was designed within the ELJABR project) are presented in Fig. 4.

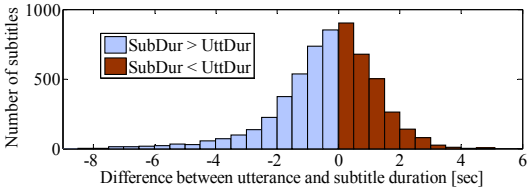

Fig. 4. Histogram: difference between subtitle durations (SubDur) and corresponding synthetised speech utterances (UttDur) for the MM synthesizer.

To illustrate the significance of the TTS system version, a new virtual male voice (denoted VM) was introduced: For the male voice (same as used in MM) the average duration of each triphone was calculated from the whole speech corpus. Then by using these values the theoretical duration of utterances for particular subtitles was determined and a similar analysis as for other synthesisers was performed.

Tab. 2. Synthesized utterances desynchronisation.

	VM	MM	MF	SM	SF
Correct begin [%]	54.6	39.4	35.7	72.3	84.7
Shifted begin [%]	45.4	60.6	64.3	27.7	15.3
Average delay [sec]	6.4	21.3	31.7	1.7	0.9

In Table 2, desynchronisation analysis is presented. The results are alarming: in the case of MUI version of our TTS system more than one half of all the utterances are delayed and the average shift is huge - more than 30 secs for MF and 20 secs for the MM synthesiser. (Detailed results for MM are presented in Fig. 5.)

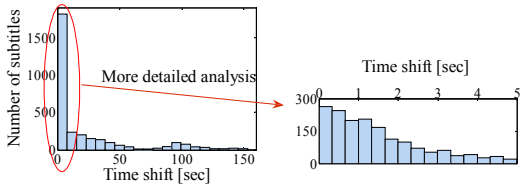

Fig. 5. Histogram: speech utterances desynchronisation for the MM synthesiser.

5. Discussion

According to the results of our experiments, the overlap of the synthesised utterances out of the given time slots constitutes a significant problem which has 4 basic solutions. However, the best result would be obtained for a combination of them.

1) Employing a faster source voice in the TTS system would simply reduce the desynchronisation problem. However all the syntesised utterances would be fast which is partly in conflict with the objective to produce a less-dynamic audio track.

2) Modification of the MUI version of our TTS system: Our experiments revealed that the MUI version selects rather longer speech unit instances during the synthesis (compare VM and MM results). A proper modification of the unit selection mechanism could speed up the resulting utterances.

3) Time-scale modification of the synthesised utterance seems to be the most suitable solution. Each utterance could be scaled individually and many of them need not be modified at all. However, a greater speech modification could introduce some degradation of speech quality and intelligibility. The analysis of necessary time scale factors for the MM synthesiser is presented in Fig. 6. (The average value is 0.79.)

4) Modification of the subtitle text is rather a theoretical solution. In some cases the subtitle text can be abridged without influence on its meaning. Primarily, it should be taken into account during the manual subtitle creation.

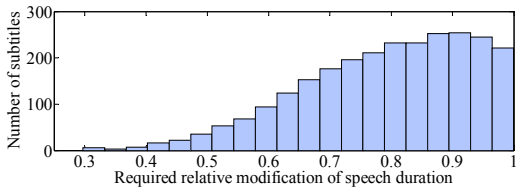

Fig. 6. Histogram: necessary time-scale factors for particular utterances (for the MM synthesizer).

6. Conclusion

The project ELJABR was introduced in this paper. The main attention was paid to the analysis of time correspondence between the automatically generated audio track and the predefined subtitles slots. It was found that more than 44 % of all subtitles available the TTS-generated synthetic speech did not fit into the predefined subtitles slots; it can cause a significant time-delay of the synthesised audio track. In other words, the current TTS system used to generate the accompanying audio track must be customised to this particular task. It is clear that great care will have to be taken in order to produce speech at faster rates especially in the cases when the generated speech overlaps the predefined subtitles time slots.

In our future work we will focus on the methods for time-scale modifications of speech, especially the WSOLA technique [7] and their integration with our TTS system. We will also discuss the amount of speech modifications, because too large amount could cause the generated audio track less intelligible. In some cases, speeding up the speech not only of the overlapping subtitles but also the speech of the neighbouring subtitles can solve the problems while modifying the speech in a reasonable extent. Both off-line (all source recordings from which the voice is built are speeded up) and on-line (only the synthesised speech which overlaps the subtitle slots is speeded up) approaches to time-scale modifications will be examined.

7. Acknowledgement

Support for this work was provided by the Ministry of Education of the Czech Rep., project No. 2C06020.

8. References

[1] A. Pražák, L. Müller, J.V. Psutka, and J. Psutka: "LIVE TV SUBTITLING - Fast 2-pass LVCSR System for Online Subtitling", *SIGMAP, LNAI 4188*, Lisbon, Portugal, 2007, pp. 139-142.

[2] J. Matoušek, D. Tihelka, and J. Romportl: "Current State of Czech Text-to-Speech System ARTIC", *TSD, LNAI 4188*, Springer, Berlin, 2006, pp. 439-446.

[3] J. Romportl: "Structural Data-driven Prosody Model for TTS Synthesis", *Speech Prosody,* Dresden, Germany, 2006, pp. 549-552.

[4] D. Tihelka, and J. Matoušek: "Unit Selection and its Relation to Symbolic Prosody: a New Approach", *Interspeech*, Pittsburgh, U.S.A., 2006, pp. 2042-2045.

[5] J. Matoušek, and J. Romportl: "Recording and Annotation of Speech Corpus for Czech Unit Selection Speech Synthesis", *TSD, LNAI 4629*, Springer, Berlin, 2007, pp. 326-333.

[6] European Broadcasting Union: "Tech.3264: Specification of the EBU Subtitling Data Exchange Format", 1991.

[7] W. Verhelst: "Overlap-add methods for time-scaling of speech", Speech Communication 30, 2000, pp. 207-221.