Automatic Dubbing of TV Programmes for the Hearing Impaired

Jindřich Matoušek, Zdeněk Hanzlíček, Daniel Tihelka and Martin Méner
Dept. of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Rep.
Email: {jmatouse, zhanzlic, dtihelka, mmener}@kky.zcu.cz

***Abstract*—This paper presents experiments with a customisation of a corpus-based unit-selection text-to-speech (TTS) system for automatic dubbing of TV programmes. The project aims at people with hearing impairments as its main goal is to produce a highly intelligible, less-dynamic, and more-undisturbed audio track for TV programmes automatically from subtitles. A two-phase synchronisation process was proposed to cope with audio-video synchronisation issues. These phases include both off-line time compression of all utterances in a source speech corpus used for TTS and on-line time compression of speech that overlaps assigned subtitle time slots. Based on a case study, in which a TTS-generated audio track of a selected movie was analysed, a simplification of to-be-desynchronised subtitle texts was proposed in order to keep time-compression factors in a reasonable extent. In this way, abrupt changes in dynamics of the produced audio track are avoided.**

I. Introduction

Nowadays, results of the research in the field of spoken language processing enable to develop applications for all — people of all ages, health conditions, native languages, and environments. Nevertheless, the most thankworthy applications from the field are undoubtedly those for people with various impairments. Among them, text-to-speech (TTS) systems are of a great importance as they enable to read text aloud. They can be exploited by people with both visual and hearing impairments. This paper presents the application of TTS technology for automatic dubbing of TV programmes with focus on hearing impaired people within 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 a broader group of TV watchers. 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 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 TV programmes are provided with subtitles (or closed captions, see Section III), 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 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 as well as 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, the alternation of very fast and normal speech, rapid changes in both voice quality and identity, 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.

A similar project was presented in [2] where a "local solution" was proposed. In that project the audio track is created from subtitles using TTS technology in a stand-alone box. Therefore, the synthetic speech is listened together with the original audio track. The project targets on people who have troubles with reading subtitle while watching a movie, i.e. on people with visual impairments and also on people with reading difficulties such as dyslexia. On the other hand, our project, introduced in [3] and further presented in this paper, proposes a "global solution" to the problems of automatic reading of subtitles. The TTS system will be used in the Czech Television, a public service broadcaster, and an alternative audio track will be delivered to ordinary home TV sets. The mixing of the TTS-generated and the original audio track will be avoided because every TV watcher will be able to choose the track according to his/her preference.

The paper is organised as follows. In Section II, the TTS technology and a TTS system used to generate the audio tracks are described. In Section III, subtitles available for experiments are specified. Section IV analyses issues related to TTS-generated speech in this very special application of TTS. In Section V, a case study, generation of synthetic audio track of a selected movie, is described. Finally, conclusions are drawn in Section VI.

II. Used technology

For automatic dubbing of TV programmes, i.e. for generation of the supplemental audio track from subtitles, text-to-speech technology was utilised. More specifically, the Czech TTS system ARTIC (Artificial Talker in Czech) [4] was

This research was supported by the Ministry of Education of the Czech Republic, project No. 2C06020, and by the University of West Bohemia, project No. SGS-2010-054. The access to the METACentrum clusters provided under the research intent MSM6383917201 is also highly appreciated.

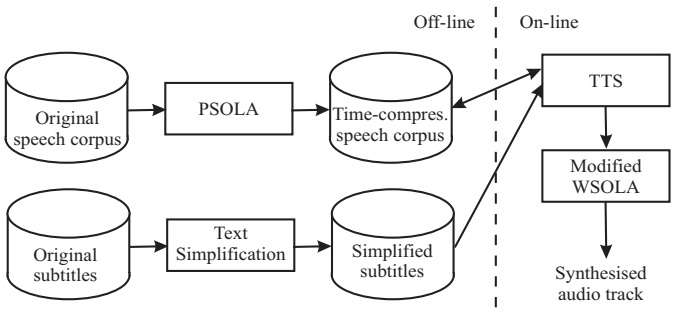

Fig. 1. A schematic overview of the system for automatic dubbing.

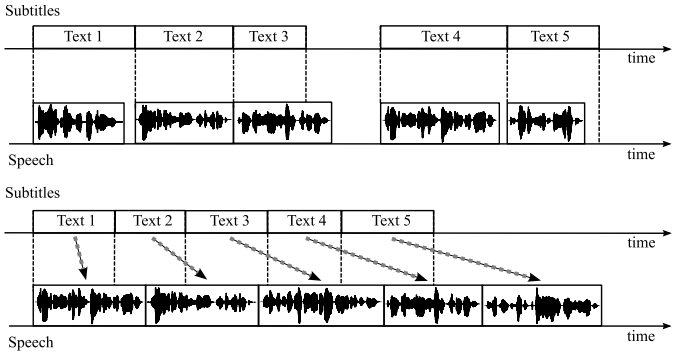

Fig. 2. Examples of synchronised (upper part) and desynchronised (lower part) synthetic speech with respect to the given time slots.

adapted. ARTIC employs a corpus–based concatenative speech synthesis method. Its principle is to smoothly concatenate (according to *join cost*) pre-recorded speech segments (extracted from natural utterances using the automatically segmented boundaries) carefully selected from a large speech corpus according to phonetic and prosodic criteria (*target cost*) imposed by the synthesised utterance. As there are usually many instances of each speech segment (mostly diphone), there is a need to select the optimal instance dynamically during synthesis run-time (using a unit selection technique).

As for the synthetic voices, two brand new voices (one male and one female) were built within the ELJABR project following the methodology described in [5]. More details about the ARTIC TTS system can be found in [4]. A schematic overview of the system for automatic dubbing is shown in Fig. 1. Each part of the system will be described further in the next sections.

III. Description of subtitles

In 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 that could be exploited when generating the supplementary 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 binary data files. Each file comprises one General Subtitle Information (GSI) block followed by a number of Text and Timing Information (TTI) blocks. In the GSI block, an overall information about the programme is defined such as original and translated title of the programme and the episode, the original language, author, and some rather technical information for both broadcasting and display. Each TTI block defines one subtitle — it is given by a subtitle text and its start time, end time, position on the screen, etc. Unfortunately, the EBU format does not currently contain any information about the characters, i.e. the assignment of subtitles to characters is not known.

For our initial analysis, a set of 20 subtitle files (5,794 subtitles, i.e. TTI blocks in sum) for various programmes was available. It comprised several documentaries, talk shows, fairy tales, cartoons, and miscellaneous movies.

IV. Main issues

As already mentioned, the supplemental audio track for TV programmes is generated automatically employing the text-to-speech (TTS) technology. A general TTS system does not usually care about speech rate. In case of a corpus-based speech synthesis system such as a unit selection one, synthetic speech tends to preserve the characteristics of speech (e.g. voice identity, speaking style, prosodic style, etc.) recorded in the source utterances. Consequently, speech rate of synthetic speech mimics the speech rate present in the source utterances, which is perfectly acceptable in applications like automatic reading of e-mails, web pages, e-books, etc., where text is usually synthesised with no requirements on a duration of the output speech utterances.

On the other hand, there are applications like automatic dubbing in which the duration of synthetic utterances does matter. In this case, when a TTS-generated synthetic utterance does not exactly fit into the time slot given by the input subtitles, serious audio-video synchronisation issues can arise.

A. Audio-video synchronisation issues

No serious problem actually arises when 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 the upper part of Fig. 2). The correspondence between subtitle and utterance starts seems to be crucial for a relaxed 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 the given subtitle time slot and overlaps into the following slot. Then, the following utterance must be delayed, and significant audio track desynchronisation occurs. In some cases (in fast dialogues), the delay accumulates, and the desynchronisation deteriorates (see the lower part of Fig. 2).

In our experiments, all available subtitles were synthesised using the current version of the ARTIC TTS system (with both male and female voice). We found that approximately 45 % of all subtitles were synthesised with duration longer

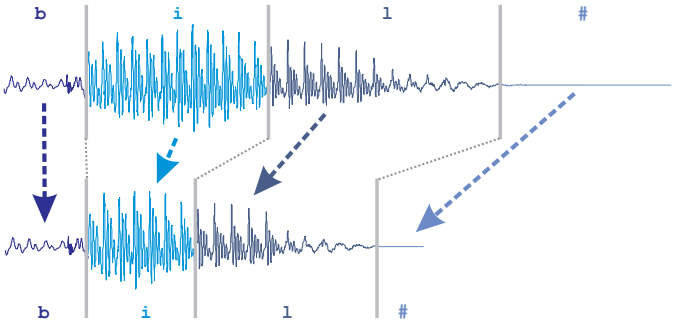

Fig. 3. An illustration of the modified WSOLA algorithm on the speech signal of the Czech word "byl" [bil] ("was" in English), # stands for a pause.

than the duration requested by the TTI block (with the average overlap 0.95 sec), possibly causing desynchronisation between the audio and video tracks. Even worse, desynchronisation accumulates in fast dialogues. Consequently, almost 61 % of all subtitles were desynchronised — the corresponding synthetic utterances do not start in the time given by the TTI block. The average delay was about 20 sec (for more details see [3]). It is obvious that desynchronisation between the TTS-generated audio track and the original video track poses a significant problem. In Section IV-B, the solution to the problem will be proposed.

B. Synchronisation of the audio track

In order to keep the audio track synchronised with the video track, standard time-scale modification techniques like WSOLA [7] can be employed to speed up the audio track. Such one-phase synchronisation runs in a subtitle-by-subtitle manner. Hence, each utterance is speeded up individually (some utterances are not modified at all), which could result in abrupt changes in dynamics of produced speech possibly causing a decrease in intelligibility of the synthesised audio track [3].

To avoid the abrupt changes in speech dynamics, synchronisation of the audio track is run in two phases. In the off-line phase, all utterances from the source speech corpus are speeded up, i.e. time-compressed by a small constant factor (0.9 for the male voice and 0.8 for the female voice) using a PSOLA-like technique [8] with no modifications of frequency speech characteristics. The PSOLA algorithm was preferred over WSOLA in this phase because PSOLA enables to recompute the positions of pitch-marks (prominent points in speech signals). In the ARTIC TTS system, the pitch-marks are used as consistent points for concatenation of voiced parts of speech [4].

In the on-line phase, each subtitle is synthesised using the time-compressed speech corpus. Duration d^{s} of the resulting utterance is then compared to the requested duration d^{r} given by the corresponding TTI block. If the duration of the synthesised utterance is longer, i.e. synthetic speech of the subtitle exceeds the assigned time slot, the utterance has to be time-compressed once more in order to fit the time slot. To do that, a modified WSOLA technique was employed. Unlike standard WSOLA, in which a single time-scale modification factor is applied to the whole speech signal [7], different phone-based time-compression factors f_p were used here. The factors f_p ($f_p = 1.0$ means no modification while e.g. $f_p = 0.5$ means a significant double time compression) were computed for each group of phones p in a sequential manner. The sequential processing ensures that duration of more variable phones like vowels, sonorants, and also pauses is compressed first up to a maximum time-compression factor $f_p^{\max}$. If the compression is not sufficient yet, the duration of other phones is sequentially compressed in the same way. Pseudocode for the computation of f_p can be written as

$phn = \{\text{pause, vowel, sonorant, fricative, affricate, plosive}\}$
$\forall p \in phn: \ f_p = 1.0$
for all $p \in phn$ **do**
  $d_p^{\mathrm{r}} = d^{\mathrm{r}} - d^{\mathrm{s}} + d_p^{\mathrm{s}}$
  if $d_p^{\mathrm{r}}/d_p^{\mathrm{s}} < f_p^{\max}$ **then**
    $f_p = f_p^{\max}$
    $d^{\mathrm{r}} = d^{\mathrm{r}} - f_p d_p^{\mathrm{s}}$
    $d^{\mathrm{s}} = d^{\mathrm{s}} - d_p^{\mathrm{s}}$
  else
    $f_p = d_p^{\mathrm{r}}/d_p^{\mathrm{s}}$
    exit
  end if
end for

where d_p^{s} is the duration of all phones from the given group p in the synthesised subtitle, and d_p^{r} is the required duration of all phones in the group p. Maximum time-compression factors $f_p^{\max}$ are currently set ad hoc with respect to the nature of each phone group p. Standard WSOLA is then employed to time-compress speech signals of phones in each phone group p with the factor f_p. An example of phone-dependent time compression of speech signals is shown in Fig. 3. A comparison of time-compression factors used during one-phase and two-phase synchronisation for a selected movie (see Section V) is depicted in Fig. 4. It can be seen that two-phase synchronisation employs less significant factors; thus, it should yield a less-dynamic audio track.

V. Case Study: Creating Audio Track for a Movie

For a case study, a comedy movie with a lot of dialogues was selected. Such a movie was preferred over documentaries and other genres because we believe that many subtitles are to be time-compressed in this movie, possibly with a significant factor. Hence, any shortcomings and speech synthesis artefacts related to the significant time-compression factors should be identified. For the purposes of this study, the EBU format was supplemented with an information about the assignment of subtitles to characters, so that we could assign male voice to male characters and female voice to female characters.

Having applied two-phase synchronisation described in Section IV-B, we found that many subtitles still required significant time-compression factors. Therefore, quality and even intelligibility of corresponding synthetic speech could be degraded. The number of subtitles affected by different time-compression factors are shown in Tab. I. For instance, it can

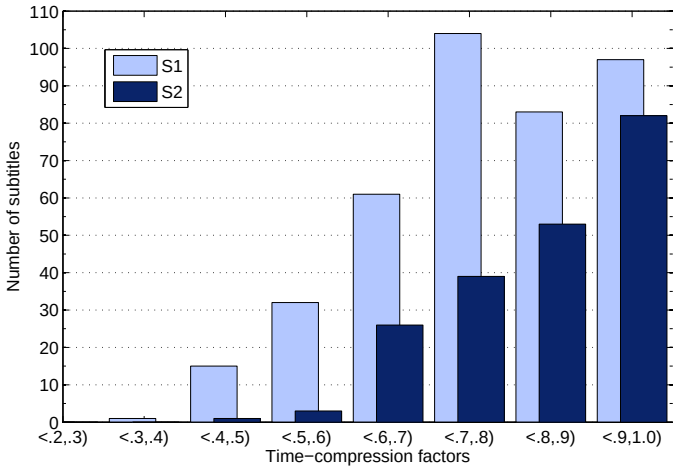

Fig. 4. Histogram of time-compression factors for one-phase (S1) and two-phase (S2) synchronisation with mean value and standard deviation 0.780 ± 0.140 for S1 and 0.841 ± 0.117 for S2 (the non-modifying factors 1.0 were excluded from the comparison — the number of subtitles not modified during on-line synchronisation was 630 for S1 and 819 for S2).

be seen that duration of 40 subtitles (3.90 % of all subtitles) was compressed with a factor less than 0.75.

It is obvious that subtitles time-compressed with significant factors would result in very fast synthetic speech, which is in conflict with our objective to produce a less-dynamic audio track. To avoid large time-scale modifications and, at the same time, to maintain the synchronicity between audio and video, we decided to simplify texts of subtitles without affecting their meaning. If we assume the time-scale modification factor 0.9 to be sufficient, 111 subtitles (almost 11 % of all subtitles) has to be simplified in our pilot movie. Ideally, the simplification of the text of subtitles should be taken into account during the process of manual preparation of subtitles (by subtitlers of the Czech Television in our case).

VI. Conclusion and future work

In this paper, we described experiments with automatic dubbing of TV programmes within the ELJABR project being solved with the co-operation with the Czech Television. We identified main issues related to the customisation of a corpus-based unit-selection TTS system to the task of automatic generation of an audio track of TV programmes. We found that about 45 % of utterances synthesised from all available subtitles exceed the assigned time slot, which caused audio-video desynchronisation of approximately 61 % of all subtitles. To keep both audio and video tracks synchronised, we proposed a two-phase procedure for time-compression of synthesised speech. Two-phase synchronisation (run both off-line and on-line) was preferred over one-phase synchronisation in order to keep changes in speech dynamics as minimal as possible. Within on-line synchronisation, a modification of the WSOLA algorithm was proposed. The modification consists in time-compression of speech in a sequential manner with respect to the nature of individual phones. In a case study, we applied the whole process of TTS-based audio track generation to a selected comedy movie. In order to keep the synchronicity between audio and video, synthetic utterances of some subtitles have to be time-compressed with a significant factor. To avoid a large compression of the duration of synthesised subtitles, a simplification of text of such subtitles was proposed.

TABLE I
The number of subtitles affected by different time-compression factors in the selected movie.

Time-compression factor	Subtitles	
	Number	Percentage
< 1.00	184	17.95
< 0.95	153	14.93
< 0.90	111	10.83
< 0.85	81	7.90
< 0.80	63	6.15
< 0.75	40	3.90

Our future work within the ELJABR project will focus on the development of a software for both the automatic detection of subtitles that will have to be simplified and the semiautomatic simplification of their texts. In order to enable a high-quality time compression of speech (even with larger factors) within the modified WSOLA algorithm, we plan to replace the expert-based setting of phone-dependent time-compression factors with a setting based on durational variability analysis of individual phones in source speech corpora. Moreover, the sequential manner of phone-dependent time compression will be replaced with a simultaneous phone-dependent compression. More movies or films in other genres are also planned to be processed to confirm the results from the case study. In case more male and female synthetic voices would be available, an algorithm for the optimal assignment of synthetic voices to characters, minimising the chance of more characters speaking with the same voice in a single dialogue scene, could be also designed.

References

[1] A. Pražák, L. Müller, J. V. Psutka, and J. Psutka, "Live TV subtitling – fast 2-pass LVCSR system for online subtitling," in *Proceedings of SIGMAP 2007*, Barcelona, Spain, 2007, pp. 139–142.

[2] S. Derbring, P. Ljunglöf, and M. Olsson, "SubTTS: Light-weight automatic reading of subtitles," in *Proceedings of NODALIDA 2009*, 2009, pp. 272–274.

[3] Z. Hanzlíček, J. Matoušek, and D. Tihelka, "Towards automatic audio track generation for Czech TV broadcasting: Initial experiments with subtitles-to-speech synthesis," in *Proceedings of ICSP 2008*, vol. 3, Beijing, China, 2008, pp. 2721–2724.

[4] 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 Artificial Intelligence. Berlin, Heidelberg: Springer, 2006, vol. 4188, pp. 439–446.

[5] ——, "Building of a speech corpus optimised for unit selection TTS synthesis," in *Proceedings of LREC 2008*, Marrakech, Morocco, 2008.

[6] "Tech.3264: Specification of the EBU subtitling data exchange format," http://tech.ebu.ch/docs/tech/tech3264.pdf, European Broadcasting Union, 1991.

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

[8] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing technique for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467.