# Dynamic Adjustment of Subtitles Using Audio Fingerprints

Lucas C. Villa Real
IBM Research
Sao Paulo - Brazil
lucasvr@br.ibm.com

Rodrigo Laiola Guimarães
IBM Research
Sao Paulo - Brazil
rlaiola@br.ibm.com

Priscilla Avegliano
IBM Research
Sao Paulo - Brazil
pba@br.ibm.com

## ABSTRACT

Anyone who ever downloaded subtitle files from the Internet has faced problems synchronizing them with the associated media files. Even with the efforts of communities on reviewing user-contributed subtitles and with mechanisms in movie players to automate the discovery of subtitles for a given media, users still face lip synchronization issues. In this work we conduct a study on several subtitle files associated with popular movies and TV series and analyze their differences. Based on that, we propose a two-phase subtitle synchronization method that annotates subtitles with audio fingerprints, which serve as synchronization anchors to the media player. Preliminary results obtained with our prototype suggest that our technique is effective and has minimal impact on the extension of subtitle formats and on media playback performance.

## Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Audio, Video*; I.7.2 [**Document and Text Processing**]: Doc. Preparation—*Format and notation, Languages and Systems, Multi/mixed media.*

## General Terms

Measurement, Experimentation, Standardization, Languages.

## Keywords

Captions; Subtitles; Timed-text; SRT; Audio fingerprinting; Synchronization.

## 1. INTRODUCTION

Captioning (or subtitling) enriches audiovisual content by providing speech information and description of representative events in a textual format. Captioning is especially important to include people with hearing loss or deafness, but its use is not limited to that domain. For instance, it is frequently the case that captions are necessary to watch a movie or TV show in a noisy environment (e.g., in an airplane) or when one is not familiar with the language or accent available in the audio streams. In fact, the benefits of captioning goes beyond the mere description of what is said or what is happening in the audiovisual content, and those benefits have been extensively reported in the research literature. The use of captioning for comprehension of a foreign language [8], vocabulary learning [6], and accessibility enhancement [5], are just some examples of these benefits.

The lifecycle of captioning can be analyzed from different perspectives. In professional post-production, extensive support is typically available; professionals create and synchronize caption data alongside the audiovisual stream. Then, the final contents are normally encapsulated in a container format and distributed on DVDs[1], Blu-Rays, broadcast television, or video on-demand services. Given that this is a very controlled workflow, major synchronization problems are not expected during consumption. Alternatively, non-professional enthusiasts are *up to the task*. This is the case particularly for online TV series and movies that are shared online. The resulting captioning files are then shared on popular online databases in a wide range of languages.

But the abundance of user-generated subtitle files comes at a price: often multiple versions are available for the same audiovisual material, leading to difficulties in identifying which of the versions is the correct one to choose. The number of different captioned files can also reflect the various editions of the media files that may or may not include advertisements, scenes from the previous or the next episode, different frame rate settings, and so on. As a result, the synchronization of the media can be compromised by choosing the *wrong* captioned file: caption entries may be displayed on the screen earlier or later than the corresponding audio and video events, as illustrated in Figure 1.

In this context, we make the following contributions. First, we perform a qualitative analysis of subtitles available in an open multi-language captioning database. This study shows that there is a great number of captioning files for a given media and that they present considerable variation on their content. Second, we introduce a two-level mechanism to dynamically adjust the presentation time of captions based not only on the regular timestamps available in captioning files, but also on audio fingerprint annotations, that work as synchronization check-points. Whenever the media player identifies such audio fingerprints, it adjusts the temporal

[1]Due to space restrictions not all technologies mentioned in this paper are Web-referenced.


*MM'15,* October 26–30, 2015, Brisbane, Australia.

ACM 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806378

```
01. ...
02. 615
03. 00:50:02,280 --> 00:50:06,046
04. TYRION: I wish I had enough poison
05. for the whole pack of you.
06.
07. 616
08. 00:50:06,120 --> 00:50:10,489
09. TYRION: I would gladly give my life
10. to watch you all swallow it.
11.
12. 617
13. 00:50:10,693 --> 00:50:12,711
14. [crowd shouting]
15. ...
```

```
01. ...
02. 729
03. 00:48:25,351 --> 00:48:28,937
04. I wish I had enough poison
05. for the whole pack of you.
06.
07. 730
08. 00:48:28,971 --> 00:48:31,139
09. I would gladly give my life
10.
11. 731
12. 00:48:31,141 --> 00:48:33,475
13. to watch you all swallow it.
14. ...
```

**Figure 1: Two distinct captioning files. Equivalent entries for the same video may differ not only in the synchronization offset (in red, left side of the '->' token), but also in the duration (in red, difference between the two) and textual content (in blue).**

offset and reschedules subsequent caption entries, in order to provide a better viewing experience. Last, we present an extension to existing subtitle file formats to support our method.

This paper is organized as follows. In Section 2 we provide an overview of related work. Next, in Section 3 we motivate our work based on a qualitative analysis of non-professional subtitle content available in a popular online database. Then, Section 4 describes our approach to address the problem of having captioning files not well synchronized with the audiovisual material. Section 5 presents the enriched subtitle format that contemplates audio fingerprint annotations. Finally, Section 6 is dedicated to concluding remarks and future work.

## 2. RELATED WORK

A large body of research has been carried out to study captioning from different perspectives [3][4][6][8][9]. For example, Hong et al. [5] propose a dynamic captioning approach that explores a set of technologies including face detection and recognition, visual saliency analysis and text-speech alignment. The authors then investigate whether subtitles placed at suitable positions can help people with hearing loss to better recognize the speaking characters. Brown et al. [1] use eye-tracking data to investigate the effect of dynamic subtitles in the viewing experience of subjects with hearing loss. Our work is related, but we focus on a different challenge: we provide a general solution to adjust the synchronization of tens (or even hundreds) of subtitle files with displaced timestamps.

From the encoding point of view, Liu and Wang [7] propose a stroke-like edge detection method for extrating captions that are hard-coded in videos. More closely related to our scenario, Bulterman et al. [2] survey many public and proprietary formats for encoding subtitles and captions. Based on a careful analysis, they describe a timed-text format that balances the need for style formatting with the requirement for more structured representation that can be easily parsed and scheduled at runtime. Our work builds on this timed-text format by proposing a second level of synchronization adjustment for captions based on audio fingerprinting.

Some well-known media players, such as VLC, are able to automatically download SubRip Text (SRT) files, one of the most popular subtitle file formats. This is possible using VLSub[2], an extension that searches for the corresponding subtitle of the loaded media based on two different approaches. In the first, the media file name is used to query a remote subtitle server. If there is a match, the corresponding subtitle file is automatically downloaded and displayed with the media. In the second approach, a hash (checksum) of the loaded media file is computed and the resulting hash is used to query a remote subtitle server. Again, if there is a match, then the corresponding subtitle file is automatically downloaded. Some of the problems with these approaches are: (1) it is easy to have file name collisions; (2) users may no longer know the original file name due to file system name mangling or to renaming; and (3) changes to the original encoding settings, such as exporting an original file in H.264 format to a QuickTime .MOV format, will alter the hash of the file. In addition, the time that the subtitle should appear and disappear on the screen might still not match. In our work, we compute the signature of audio events alone. These are not likely to change drastically when the original audio is re-encoded with different settings. Thus, we propose the usage of an audio fingerprint to univocally determine the moment a check-point is achieved. Once this audio fingerprint is identified on the file, the temporal offset of all the subtitles entries are dynamically adjusted to the specific media file.

## 3. QUALITATIVE ANALYSIS

To get an understanding of the quality and differences among subtitle files available on the Internet, we picked a Top-5 movie ranked by IMDb and analyzed all of the corresponding subtitle files available for such movie at the OpenSubtitles.org[3] repository. Since movies are different from TV series in some ways (for instance, a TV series episode may start with scenes from the previous episode, and finish with a preview of what is coming in the next one), we also analyzed the subtitles of an episode of a popular TV series (according to IMDb).

For the movie, there were 193 subtitle files distributed across 33 different languages. The languages with the highest number of captioned files are English (35), followed by Brazilian Portuguese (18), and Spanish (16). Since these files often come with advertisements and credits that are introduced as actual caption entries[4], all files were sanitized

[2]https://github.com/exebetche/vlsub

[3]We used OpenSubTitles.org because it offers an open API

[4]Interestingly enough, it is not uncommon to find the very

**Table 1: Analysis of captions for a popular movie. The duration is given by the difference between the last and first caption entries.**

| Language | Time at first caption | Total duration |
|---|---|---|
| English | 26 ± 5 s | 8781 ± 191 s |
| Brazilian Portuguese | 14 ± 3 s | 8960 ± 575 s |
| Spanish | 15 ± 5 s | 8779 ± 341 s |
| French | 16 ± 12 s | 8730 ± 275 s |
| Turkish | 14 ± 7 s | 8762 ± 184 s |

before being properly analyzed.

Table 1 shows there are variations in the starting time and total duration covered by the captions even within the same language. The meta-data[5] of the movies associated with these subtitle files indicate that sometimes their duration also differs, since some are extracted from DVDs and Blu-Rays and some others are recorded from broadcast TV. Second, at times there is an offset between translated subtitles and subtitles in the original language (English). We found that this comes from audio differences in dubbed versions of the movie. The standard deviation is mostly caused by the use of audio-descriptions (i.e., caption entries describing audio events such as an opening door) and different reference times (e.g., some subtitle files have their first caption entry starting at time zero, even though the actual audio event happens only a few seconds from the start).

In our analysis of one episode of a popular TV series, a total of 170 subtitle files were collected, covering 38 languages. We inspected patterns from Brazilian Portuguese (19 subtitle files), English (13 files), and Spanish (11 files). In contrast to the movie file, the presence and absence of prologues and epilogues was often observed and caused a large standard deviation at the time of presentation of the first caption (45 seconds for an average of 88 seconds in Brazilian Portuguese). This is more clearly seen in Figure 2, which depicts the presentation time of caption entries of files with and without a prologue. Surprisingly, we have also noticed subtitle files with differing *pace*. That is, even when they are properly aligned at the first caption entry, they eventually get out of sync with each other. Our analysis found that the associated video files had different encoding settings (frame rate) which caused one of them to be played back quicker. This finding is shown in Figure 3.

This study reinforces that finding a subtitle file that will fit well with a given media can be a challenging task. The variations in the starting times and the problems caused by different encoding settings support our premise that a mechanism to automate the synchronization of such a large range of potential subtitle files is required.

## 4. OUR APPROACH

The golden standard for captioning is a file containing, essentially, the textual entries to be displayed (thereafter called subtitles) and the moment of that presentation. The media player, when playing back a multimedia file/stream, renders also the textual entries specified on the subtitle file for that moment.

same subtitles disguised in different file names but with different names listed in the credits. This suggests that ownership infringement could be a recurrent problem in subtitle sharing communities on the Internet.

[5]Information such as frame rate, resolution, and bit rate of the multimedia file.

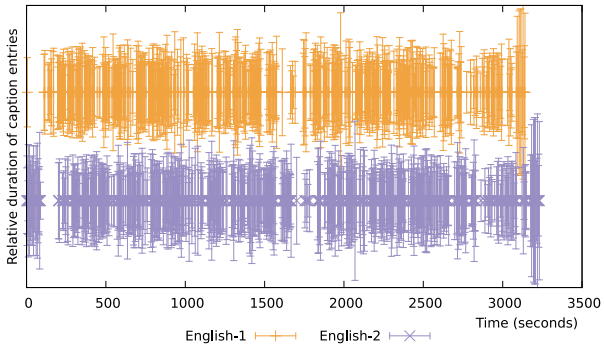


**Figure 2: Caption entries for an episode of a TV series. The bottommost captions include a prologue.**

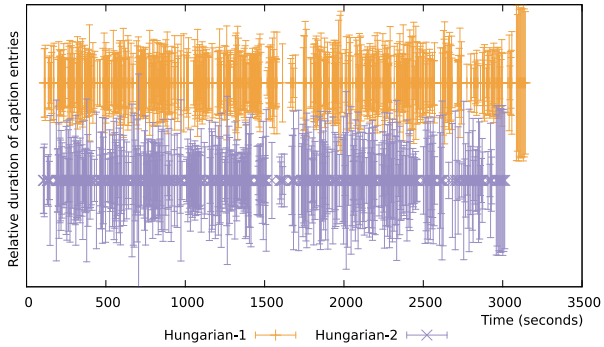


**Figure 3: Caption entries for two videos encoded with different frame rates: differences begin small but increase over time (represented by the X axis).**

Our method to dynamically synchronize a subtitle file with media streams works on two steps: (1) on the creation or modification of a subtitle file through an authoring software and (2) on the playback of that subtitle file along with its media streams.

The first step consists of electing subtitle entries as synchronization anchors. Ideally, each subtitle file has at least two anchors, chosen, preferably, from moments close to the begining and the ending of the file. The authoring software then takes short audio fingerprints that describe the main audio component of these anchors. In other words, it extracts the *signature* of the audio signal associated with each anchor. Next, all the fingerprints are added to the subtitle file with their corresponding offset time.

The second step of the synchronization happens on the playback of the media file. Essentially, it begins with the opening of the media streams and with the parsing of the subtitle file, from where the caption entries and the fingerprints are read. The media player seeks to the offsets of the first and last anchors and calculates the audio fingerprint of the media streams at these offsets. Subsequently, they are compared against the fingerprints specified on the subtitle file. The two media files are considered in perfect sync if the calculated and the reference fingerprints match.

If the fingerprints do not match, then the media player seeks to the beginning of the media file and computes its audio fingerprints until there is a match with the signature of the first anchor annotated on the subtitle file. The difference between the media player's time offset and the timestamp

of the matched anchor is then propagated to all subsequent caption entries. A comparison between the last anchor's fingerprint is then performed to assure the correct synchronization. If discrepancies are still found, the media player moves back to the first anchor point and starts to search for the second anchor point, in order to correct the subsequent subtitles one more time. This process is repeated until the fingerprint of the last anchor matches the one from the file. Finally, the media player utilizes the updated caption timestamps to synchronize the playback.

A special case applies to non-seekable media streams (as is the case of broadcast and streaming services.) Since there is no indication of the beginning of the stream, the media player needs to continuously compute the audio fingerprints until a match is found. At that point in time, the subtitles can be put in sync with the media streams and the media player can start rendering them. This process can be performed at uniform intervals, to guarantee the synchronization throughout the playback.

# 5. ENRICHED SUBTITLE FORMAT

The method proposed in this paper has been prototyped with the VLC media player and Chromaprint[6], a library that implements a custom algorithm for extracting fingerprints from audio sources. Besides the purpose of validating the core idea, the prototype also helped us better understand the performance impacts of the fingerprinting process and to investigate how to integrate audio signatures into existing subtitle file formats.

Our prototype includes support for the SRT format, depicted in Figure 4, which is comprised of three elements: (1) the sequence number of the caption entry (lines 01 and 05), (2) the starting and ending times in which the caption entry must be presented on the screen (lines 02 and 06), and (3) the textual entries. To overcome the lack of support for including meta-data in SRT, we chose to represent fingerprints as caption entries with zero duration. The string `@fingerprint@` indicates the presence of the audio signature computed by the fingerprint library, as shown in Figure 4. Other subtitle file formats could embed the audio fingerprints as proper meta-data and a binary format would store the fingerprints more efficiently. Given the popularity of SRT, though, it is essential to include new features while maintaining backwards compatibility with that format.

```
01. 0
02. 00:00:00,000 --> 00:00:00,000
03. @fingerprint@ -177,-168,-167,...,934,934,904
04.
05. 1
06. 00:01:46,144 --> 00:01:48,831
07. - ( woman laughing )
```

**Figure 4: Sample fingerprint anchored at time zero.**

In our method, anchors do not necessarily need to be associated with actual captions or subtitles. It is perfectly possible to compute audio fingerprints from the very first seconds of the multimedia streams and annotate them as an empty caption entry, as shown in Figure 4 (line 02). By doing so, a minimal amount of fingerprint processing can be performed during playback.

[6]http://acoustid.org/chromaprint

# 6. CONCLUSION

In this paper we presented three contributions to the area. First, through a qualitative analysis of publicly available subtitles we identified sources of synchronization problems between captioned files and their corresponding media. Second, we developed a method to dynamically sychronize subtitles based on audio fingerprints. Third, we proposed an extension to the subtitle file formats currently in use so they can benefit from our dynamic synchronization method.

As future work, we intend to deepen the investigation on the causes of the discrepancies between subtitle files, in order to propose new or enhance our extension on subtitle files. Moreover, we would like to conduct an experiment to measure the sensitivity of users to subtitle's dyssynchrony. This information will be useful for the calculation of the optimized frequency of anchor points in subtitle files.

# 7. ACKNOWLEDGMENTS

This work has been partially funded by Brazilian's FINEP / MCTI under contract no. 03.11.0371.00.

# 8. REFERENCES

[1] A. Brown, R. Jones, et al. Dynamic subtitles: the user experience. In *Proceedings of the 2015 ACM International Conference on Interactive Experiences for TV and Online Video*. ACM, 2015.

[2] D. C. Bulterman, A. Jansen, et al. An efficient, streamable text format for multimedia captions and subtitles. In *Proceedings of the 2007 ACM symposium on Document engineering*, pages 101–110. ACM, 2007.

[3] C. Concolato and J. Le Feuvre. Live http streaming of video and subtitles within a browser. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 146–150. ACM, 2013.

[4] M. Federico and M. Furini. An automatic caption alignment mechanism for off-the-shelf speech recognition technologies. *Multimedia tools and applications*, 72(1):21–40, 2014.

[5] R. Hong, M. Wang, et al. Dynamic captioning: video accessibility enhancement for hearing impairment. In *Proceedings of the international conference on Multimedia*, pages 421–430. ACM, 2010.

[6] G. Kovacs and R. C. Miller. Smart subtitles for vocabulary learning. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 853–862. ACM, 2014.

[7] X. Liu and W. Wang. Robustly extracting captions in videos based on stroke-like edges and spatio-temporal analysis. *Multimedia, IEEE Transactions on*, 14(2):482–489, 2012.

[8] K. Rooney. The impact of keyword caption ratio on foreign language listening comprehension. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, 4(2):11–28, 2014.

[9] J. Tiedemann. Synchronizing translated movie subtitles. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'2008)*, 2008.