IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 24, 2021, accepted September 29, 2021, date of publication October 8, 2021, date of current version October 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3119201

Automatic Subtitle Synchronization and Positioning System Dedicated to Deaf and Hearing Impaired People

BOGDAN MOCANU[1], (Member, IEEE), AND RUXANDRA TAPU[1,2], (Member, IEEE)

[1]Telecommunication Department, Faculty of ETTI, University Politehnica of Bucharest, 060042 Bucharest, Romania
[2]ARTEMIS Department, Institut Polytechnique de Paris, Télécom SudParis, Laboratoire SAMOVAR, 91000 Évry, France

Corresponding author: Ruxandra Tapu (ruxandra.tapu@telecom-sudparis.eu)

This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS - UEFISCDI under Project PN-III-P1-1.1-TE-2019-0420, within PNCDI III.

ABSTRACT In this paper, we introduce a subtitle synchronization and positioning system designed to increase the accessibility of deaf and hearing impaired people to multimedia documents. The main contributions of the paper concern: a novel synchronization algorithm able to robustly align, without any human intervention, the closed caption with the audio transcript and a timestamp refinement technique that adjusts the subtitle segments duration with respect to the audiovisual recommendations. Finally, we introduce a novel method that performs a high level understanding of the multimedia content, in order to determine the subtitle optimal positions, within the video frame, such that they do not overlap with other relevant textual information. The experimental evaluation performed on a large dataset of 30 videos taken from the French national television validates the approach with average accuracy scores superior to 90% regardless on the video genre. The subjective evaluation of the proposed subtitle synchronization and positioning system, performed with actual hearing impaired people, demonstrates the effectiveness of our approach.

INDEX TERMS Subtitle/closed caption synchronization, audiovisual recommendations, anchor words, tokens, subtitle positioning.

I. INTRODUCTION

In the last years, the volume of available live programs transmitted over Internet has shown an exponential growth. In order to facilitate the access to information, most European TV broadcasters transmit and distribute, together with the audio and video signals, textual information. Most often, such information is presented under the form of video subtitles or closed captions. Even though the subtitle/ closed caption does not match exactly the audio (speech) transcription and corresponds rarely to the exact word utterance, the associated text document (developed by a human transcriber) captures all the semantic value according to the available visual space and slot of time [1].

The subtitles are provided in order to convey the content of foreign language dialogue and also to meet the needs of a significant number of deaf/hearing impaired people. In the recent statistics published by the World Health Organization it is stipulated that for people over 50 years old the hearing impairments are becoming progressively common and by the year of 2050 over 900 million people will suffer from hearing loss [2]. For such people, the subtitles represent the most efficient way to access the audio-video content of a TV program. For this reason, the TV regulations, applied worldwide in a majority of countries, compel the TV content providers to offer subtitles/closed captions for almost all broadcasted programs.

The problem becomes critical in the case of live scenarios. Here, a subtitling environment is set up in the studio, where a regular human develops on-the-fly the audio transcript. The associated textual content is delivered several seconds after the speech fragment. For this reason, most of the live-transmitted TV programs, present a significant delay (up to 30 seconds) between the audio stream and the moment in time when the corresponding subtitle is ready to be displayed on the screen. As a consequence, the subtitle/closed caption delivered to the user is unsynchronized. Moreover, the delay between the textual and visual information may be different

The associate editor coordinating the review of this manuscript and approving it for publication was Arif Ur Rahman.

for various parts of the subtitle. This effect is highly disturbing and can strongly affect the user comprehension.

The only solution to overcome this limitation, largely adopted by TV broadcasters, is obtained with the help of program replay. In this case, as the subtitle is available in advance, a temporal re-alignment process is manually performed by human specialists [3]. However, such a solution is time and money consuming.

In this paper, we introduce a novel completely automatic subtitle synchronization methodology dedicated to replay videos, which notably aims at ensuring the coherence between the audio stream and the textual information displayed on the user screen. The main contributions of the paper are the following:

(1). A robust, automatic method for synchronizing subtitle/closed captions with the audio channel of a video file. The system receives as input the closed caption data, produced by human transcribers. This data contains useful information, but suffers from severe misalignments with the audio channel, because of the manual production process involved. By using publically available automatic speech recognition (ASR) softwares we are able to generate an approximate transcription of the parsed audio channel. The words of the ASR transcript have a relatively low degree of accuracy and are therefore incomplete. Also, the performances of the ASR systems are highly dependent on the language model. However, they offer the advantage of being perfectly synchronized with the content. The text alignment algorithm introduced in this paper makes it possible to robustly align, without any human intervention, the closed caption with the audio transcript and thus to synchronize the caption with the content.

(2). A refinement technique that adjusts the subtitle durations such that they respect the audiovisual recommendations specified in France by the CSA (*Conseil Supérieur de l'Audio-visuel*).

(3). In addition, to the best of our knowledge, we introduce the first framework that performs a high level of understanding of the video content in order to position the subtitle within the video frames such that they do not overlap other textual information that can be present in the scene.

The rest of the paper is organized as follows. Section II reviews the speech emotion recognition state-of-the-art techniques. Section III describes the proposed architecture, and details the key elements involved. Section IV presents the experimental setup and the evaluation results obtained. Finally, Section V concludes the paper and opens some perspectives of future work.

II. STATE OF THE ART REVIEW

The state of the art techniques that tackle the issue of audio/subtitle alignment are mostly based on automatic speech recognition (ASR) systems, previously trained on various datasets and language models in combination with a video decoder, in order to adjust the timestamps of the video subtitle. The main difficulty encountered when building a speech recognizer in a new target language (*i.e.*, French, Spanish, Portuguese...) is related to the availability of manually transcripts and aligned training data.

One of the first papers addressing the issue of aligning very long audio files to their corresponding textual transcripts is introduced in [4]. The key principle of the proposed algorithm relies on a speech recognition procedure that gradually restricts the dictionary and the language model. In addition, the authors introduce the concept of islands of confidence that are likely to be aligned correctly based on an acoustic confidence metric. An iterative procedure is then applied in order to successively refine/optimize the alignment between anchors.

The open source software toolkit so-called SailAlign [5] aims at providing robust speech to text synchronization. First, the audio stream is segmented into small chunks. To avoid cutting a word into two parts, a voice activity detection module is employed. Then, a speech recognition algorithm is applied in order to identify the lexical content of each individual speech segments. The regions that can be reliably aligned are identified by applying the so-called minimum number of words criterion, while the rest of the audio is considered as unaligned.

Different approaches such as those introduced in [6], [7] start by training a biased language model (LM) on textual information together with ASR. The resulting transcript is aligned to the reference text by using dynamic programming.

The works described in [8]–[10] show that some sort of text alignment is possible when very little information about the language is available. Thus, in [8], inaccurate transcripts generated from lectures and presentation videos are aligned to the audio stream using a small set of acoustic units, detected heuristically as anchor points. In [9], Hidden Markov Models (HMM) are used in order to perform a force alignment at the sentence level by using a subset of short utterance trained with data from several languages. By using a set of generalized models computed on phonetic groups, the system can reliable align text to speech while being robust against the transcription errors. Finally, in [10], by exploiting the syllable information in speech and text transcripts the authors introduce a technique for text alignment at the sentence level.

Another family of approaches is focused on efficient ways of training the speech recognition frameworks in order to increase the system robustness. Such methods include: the adaptation of existent ASR systems, trained on the same language, but with a different dialect [11], cross-language bootstrapping [12] or training seed acoustic models from very small volumes of manually aligned data [7].

An ASR system trained from scratch using approximately aligned text transcripts to the audio recordings is introduced in [13]. The training data (*e.g.*, audio books, parliamentary speeches) have been harvested from the Internet. First, the phoneme sequences are aligned to the normalized text transcripts through dynamic programming. Then, the correspondence between the phonemes and the graphemes is obtained through a matrix of approximation. Finally, the audio data is split into short segments and the ASR is trained using the

Kaldi toolkit [14]. In [15], a system for building graphemes for any language written in unicode is proposed. The main idea is to extract the attribute for the graphemes automatically using the features from the unicode character description.

From the methods presented above it can be observed that in most of the cases the synchronization of the audio tracks with the text transcripts is typically solved by forced alignment performed at the phonetic level. However, such approaches show quickly their limitations when dealing with very long audio tracks or acoustically inaccurate text transcripts. In such cases, a more sophisticated analysis is required. To this purpose, in [16], the authors propose using probabilistic kernels (similarity functions) about the speaker behavior in order to improve the text alignment. Recently, in [17] in order to handle the variability of amateur reading and improve the performance of text alignment systems, the authors introduce a human in the loop approach. They propose to use a finger text tracker in order to implement a metaphor (anchor points). Then, the metaphors are recorded during the collocated, read-aloud session using tablet e-readers.

The introduction of deep learning algorithms in the context of ASR, have improved the speech system performance and robustness [16], [18]–[21]. While the increase in accuracy is significant, the deep learning still plays a limited role in the traditional pipelines [22]. Nowadays, the challenge is to perform the speech to text transcript when dealing with noisy environmental conditions, with background clutter or when multiple actors are speaking in the same time.

However, as indicated above, the link between the subtitle and the locution is a case of a classical issue: the alignment of symbol sequences. Even though the ASR can potentially return accurate synchronization between audio and text, reliable ASR for all portions of the video stream is a difficult task mainly due to high variability of the spoken environments, speakers and speech types. Spoken environments can vary from clean settings (*e.g.,* studio recordings) to noisy scenes (*e.g.,* outdoor recordings, speech mixed with background music or effects). The reader speaking may be classified into two major classes: dictation (*e.g.*, the newsreader presenter) and spontaneous (*e.g.,* interviews taken on the street, debates or talk-shows). In addition, the ASR technology needs to deal with uncontrolled vocabulary (especially for movies or TV programs). Any system dedicated to robust subtitle/caption alignment should carefully take into account such issues. A second challenge that arises when performing the synchronization is to determine what technique to apply when the text transcript does not match exactly the spoken words of the media. In most of the cases, when a media creator edits the transcripts it removes the grammatical errors or other types of extraneous words spoken repeatedly during the course of the audio stream (*e.g.*, bref, voila, uh, aha...). Even more challenging are the cases when the person's speech frequency is very high and the transcriber cannot follow all transmitted information. In this case, the captioner reformulates the idea in fewer phrases that are not perfectly correlated with the audio stream. The third challenge that needs to be overcome is the time necessary to perform the alignment. In addition, the cost of such systems should not be excessive for the television companies.

In this paper, we introduce a novel methodology for automatic synchronization of the captions/subtitles with the audio-video streams. Compared with other approaches of the state of the art, we do not focus our attention on the ASR training or on the data acquisition, but on efficient ways to use anchor words in order to position the text transcript in noisy scenes or with multiple speaking persons. In contrast with state of the art frameworks, the proposed system respects the audio-visual recommendations expressed by the CSA (Conseil Supérieur de l'Audiovisuel) (*i.e.*, Rec. "*Charte relative á la qualité du sous-titrage á destination des personnes sourdes ou malentendantes*" [23]) regarding the appropriate time of reading when a phrase is displayed on the screen.

An additional feature of the proposed method concerns the spatial positioning of the subtitle in the video frames. To the best of our knowledge, the proposed system is the first one that automatically adjusts the position of the subtitle based on the visual content. This makes it possible to avoid displaying subtitles on particular areas of interest that can typically include useful video text information.

III. PROPOSED METHOD

In the following section, we will refer to two types of text documents that are supposed to be available for a same TV program [24]: (1) *closed captions (CC)* - which are program transcripts, prepared by a human transcriber, that contain a time code for every group of 3-10 words and may or not be prepared in real-time; (2) *audio speech recognition (ASR) outputs* – which denote text generated automatically by a speech recognition software. In this latter case a time code is associated to each word. The objective is to automatically align the closed caption with the transcript of the ASR output. First, we have carried out an analysis of both the closed caption and the automatic speech recognition outputs. We have observed that various types of errors can occur. For the CC, four major types of errors can appear:

(1). spelling errors: corresponds to words that were missed, added or misspelled by the captioner;

(2). interpretation errors: the human transcriber paraphrases or replaces the spoken word/words with a similar one;

(3). group errors: represent block of words or phrases that are skipped in the closed caption text from the program transcript;

(4) repeated words or lines: correspond to words that are typed twice or multiple times by mistake.

In the ASR output three types of errors have been identified:

(1). homophone substitution errors: represent errors generated by the ASR system that substitutes a correct word to its phonemes rather the complete word;

(2). *misinterpretation errors*: the ASR replaces a word with a different one;

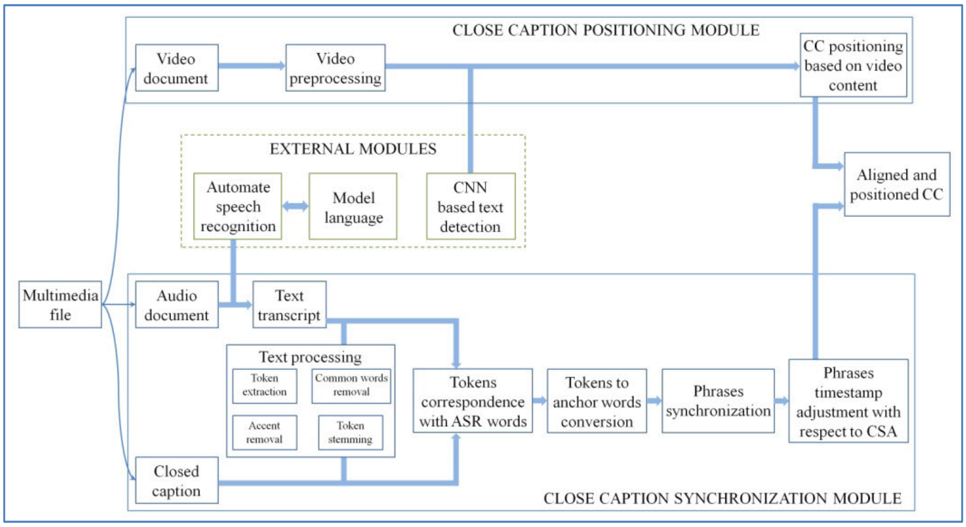
CLOSE CAPTION POSITIONING MODULE
Video document
Video preprocessing
CC positioning based on video content
EXTERNAL MODULES
Automate speech recognition
Model language
CNN based text detection
Aligned and positioned CC
Multimedia file
Audio document
Text transcript
Text processing
Token extraction
Common words removal
Accent removal
Token stemming
Tokens correspondence with ASR words
Tokens to anchor words conversion
Phrases synchronization
Phrases timestamp adjustment with respect to CSA
Closed caption
CLOSE CAPTION SYNCHRONIZATION MODULE

FIGURE 1. The proposed CC/ASR synchronization and positioning system: workflow and modules involved.

(3). *group errors*: represent a set of words (block of words) that are missed entirely from the audio stream by the ASR.

Dealing with these various types of errors for ensuring a robust CC/ASR alignment is the key challenge that we attempt to solve with the proposed method. The proposed CC/ASR alignment architecture is presented in Figure 1. The system takes as input a multimedia file that contains the audio and video streams that are supposed to be perfectly synchronized.

Each multimedia file has a starting time (ST) and an end time (ET), which define the timeline for the considered document. To each multimedia file, a closed caption document is also associated that includes the text transcript of the corresponding audio information.

The system exploits three external components: a language model toolkit, an automatic speech recognition system and a text detection framework (in our case based on convolutional neural networks). The speech recognizer adopts the French language model, takes as input the audio stream of the target video and generates as output transcriptions and time codes for the identified words. The proposed alignment algorithm reassigns a new timestamp for each phrase in closed caption. The text detection framework is designed to identify the position of existing text within the video stream. This information makes it possible to adjust the position of the displayed subtitle such that avoiding overlaps with the visual content.

Concerning the ASR system, for the time being we have adopted the open-source CMU Sphinx system [25], which offers a fair compromise between the system performance and required memory/computational resources. Thus, as shown in the comparative evaluation of ASR systems presented in [26], carried out on a set of 30 news videos of about 30 min length, the word error rate (WER) achieved by Sphinx is of 35%, which is relatively high. However, for the words that are correctly identified, the ASR is able to accurately capture the timestamps of the audio text transcript. In addition, Sphinx provides an already trained French language model. We can expect that by adopting a more sophisticated/robust ASR system, such as one of the emerging CNN-based approaches [27], the overall performances to increase.

Let us now detail the synchronization algorithm proposed, which represents the core of our method.

A. SUBTITLE/CLOSED CAPTION SYNCHRONIZATION ALGORITHM

The synchronization algorithm takes as input the closed caption manually generated by the transcriber and the text transcript, provided by the ASR module. The output consists of a set of timestamps associated to the closed caption phrases.

The first step of our algorithm concerns the pre-processing of the input texts.

1) PRE-PROCESSING AND INITIALIZATION

The objective is to convert the closed caption and the ASR output into a standard form, in order to increase the matching probability. First, both transcripts are divided into atomic units of text known as tokens, where each token is delimited by a white space. Then, each token is processed to remove all punctuation and non-alphabetic characters. Uppercase words are also converted to lowercase words.

Words with less than four characters are removed from both text transcripts. In addition, in order to obtain a normalized text (words with no accents or diacritics, plural words…) we apply text stemming. Very common words (*e.g.*, voila, chose, comme, parce que…) are also removed in order to increase the discriminating power of the extracted tokens. Figure 2 presents the various steps involved in the text

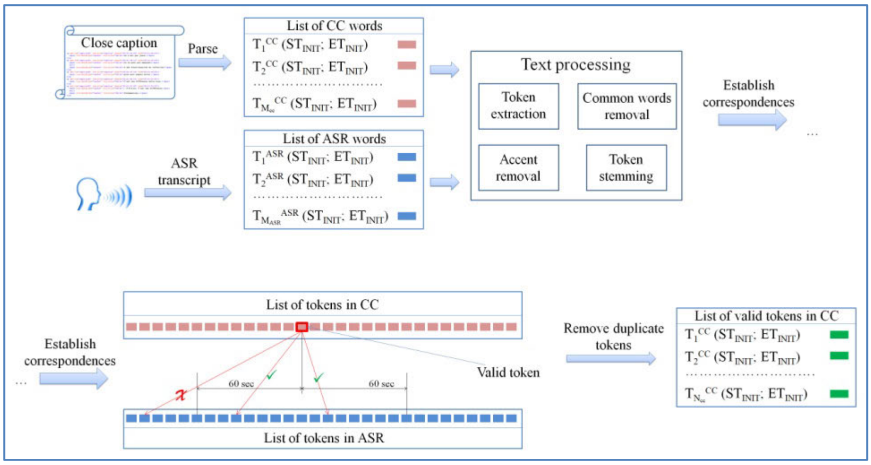
Close caption
Parse
List of CC words
T_1^{CC} (ST_{INIT}; ET_{INIT})
T_2^{CC} (ST_{INIT}; ET_{INIT})
$T_{M_{cc}}^{CC}$ (ST_{INIT}; ET_{INIT})
ASR transcript
List of ASR words
T_1^{ASR} (ST_{INIT}; ET_{INIT})
T_2^{ASR} (ST_{INIT}; ET_{INIT})
$T_{M_{ASR}}^{ASR}$ (ST_{INIT}; ET_{INIT})
Text processing
Token extraction
Common words removal
Accent removal
Token stemming
Establish correspondences
...
...
Establish correspondences
List of tokens in CC
60 sec
60 sec
Valid token
List of tokens in ASR
Remove duplicate tokens
List of valid tokens in CC
T_1^{CC} (ST_{INIT}; ET_{INIT})
T_2^{CC} (ST_{INIT}; ET_{INIT})
$T_{N_{cc}}^{CC}$ (ST_{INIT}; ET_{INIT})

FIGURE 2. Text pre-processing and token matching between CC and ASR channels.

processing stage. Let us recall that a time-code is associated to each individual word in the ASR transcript.

At the end of the pre-processing phase, we obtain two token sequences, respectively denoted by $\{T_i^{CC}\}_{i=1,\ldots,N_{CC}}$ and $\left\{T_j^{ASR}\right\}_{j=1,\ldots,N_{ASR}}$ for the CC and ASR streams, with N_{CC} (resp. N_{ASR}) being the total number of tokens of the CC (resp. ASR) streams. Each token in the ASR stream has an associated timestamp that provides its precise temporal localization with corresponding start time (ST) and end time (ET). However, this is not the case of CC tokens. Here, the human transcriber provides solely the timestamps associated to entire phrases, with a phrase start time (ST_{P_I}) and end time (ET_{P_I}).

In order to be able to estimate an initial timestamps to the CC tokens that are present in the considered phrase, we apply the following procedure. Based on the number of characters within the sentence/phrase (No_{C_P}) we estimate the initial token word temporal elements (ST_{INIT} and ET_{INIT}) using the following equations:

$$Time\_per\_ch_{INIT} = \frac{ET_{P\_I} - ST_{P\_I}}{No_{C\_P}}, \quad (1)$$

$$ST_{INIT} = ST_{P\_I} + Time\_per\_ch_{INIT} * No_{C\_TO\_T}, \quad (2)$$

$$ET_{INIT} = ST_{INIT} + Time\_per\_ch_{INIT} * No_{C\_OF\_T}, \quad (3)$$

where $No_{C\_TO\_T}$ represents the number of characters covered, within the phrase, until the considered token, while $No_{C\_OF\_T}$ denote the number of characters of the current token.

Let us observe that both parameters $No_{C\_TO\_T}$ and $No_{C\_OF\_T}$ are obtained from the initial CC transcript, before any pre-processing steps.

2) ANCHOR WORD IDENTIFICATION AND TOKEN MATCHING ALGORITHM

After the pre-processing step, the objective of the alignment algorithm is to determine for each token in the closed caption, the corresponding token in the ASR output. This makes it possible to assign to each CC token the timestamp of the corresponding ASR token. The anchor words are defined as CC tokens for which a matching with a token from the ASR stream has been established. The key challenge is to determine such anchor words in a reliable manner. This is equivalent to setting up a robust token matching algorithm. In this section, we describe the token matching procedure proposed.

The word alignment algorithm is processing one by one the closed caption tokens, in order to determine for each of them the corresponding token in the ASR output. For each CC token, a forward/backward search procedure is applied. A maximum time shift on each direction of searching (*i.e.,* forward and backward) is admitted, defining thus a temporal search window. The typical values that we have considered for the forward/backward time shift can vary between 30 and 90 seconds, but this parameter can be tuned for various types of content. When a unique match is identified in the considered search interval, the token is considered as an *anchor word*. If multiple matches are determined for the considered CC token, a further analysis is performed.

Let us denote by T_c^{CC} the current token under analysis in the CC stream and by $\mathcal{N}\left(T_c^{CC}\right) = \left\{T_i^{CC}\right\}_{i=c-N, i\neq c}^{c+N}$ the set of its neighboring tokens, with N being the number of backward and forward neighboring tokens. In our work, we have set the value of N to 5, in order to avoid increasing the computational complexity.

For each CC token in the $\mathcal{N}\left(T_c^{CC}\right)$ set, we examine its correspondences determined in the ASR stream, as described in the following. Let us denote by $\aleph^{ASR}\left(T_c^{CC}\right)$ the set of tokens in the ASR stream that are matched with the current T_c^{CC} token. In a similar manner, let $\aleph^{ASR}\left(T_i^{CC}\right)$ denote the set of ASR tokens that are matched with each neighboring token $T_i^{CC} \in \mathcal{N}\left(T_c^{CC}\right)$.

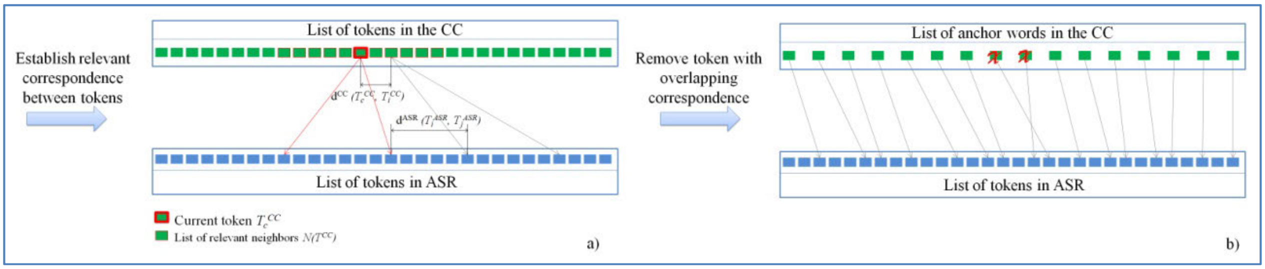
Establish relevant correspondence between tokens
List of tokens in the CC
$d^{CC}(T_c^{CC}, T_i^{CC})$
$d^{ASR}(T_i^{ASR}, T_j^{ASR})$
List of tokens in ASR
Current token T_c^{CC}
List of relevant neighbors $N(T^{CC})$
a)
Remove token with overlapping correspondence
List of anchor words in the CC
List of tokens in ASR
b)

FIGURE 3. Token assignment to the anchor word list: (a). Token conversion to anchor words; (b). Discarding temporally conflicting anchor words.

For each pair formed by the current token T_c^{CC} and one of its neighbors $T_i^{CC} \in \mathcal{N}\left(T_c^{CC}\right)$, we compute two temporal distance values. A first one, denoted by $d^{CC}(T_c^{CC}, T_i^{CC})$, is simply the temporal distance between the two considered tokens, with respect to their corresponding timestamps in the CC stream. For each token, a timestamp is defined as a starting point (ST) and an end point (ET). The temporal distance of two tokens is defined in our case as the absolute difference between the corresponding starting times. The second distance, denoted by $d^{ASR}(T_c^{CC}, T_i^{CC})$, is computed based on corresponding tokens in the ASR stream and is defined as described in the following equation:

$$d^{ASR}\left(T_c^{CC}, T_i^{CC}\right) = \min_{T_i^{ASR} \in \aleph^{ASR}(T_C^{CC})} \times \min_{T_j^{ASR} \in \aleph^{ASR}(T_i^{CC}} d^{ASR}(T_i^{ASR}, T_j^{ASR}) \tag{4}$$

With the help of the above defined distances, the current T_c^{CC} token is declared as an anchor word if and only if the following condition is satisfied:

$$\forall T_i^{CC} \in \mathcal{N}\left(T_c^{CC}\right), \quad d^{ASR}\left(T_c^{CC}, T_i^{CC}\right) \leq \alpha \cdot d^{CC}\left(T_c^{CC}, T_i^{CC}\right), \tag{5}$$

where α is a predefined parameter. In our experiments, a value of α between [2]–[4] returned similar results.

The procedure of anchor words identification is illustrated in Figure 3a. Let us observe that, in some cases, the temporal order of the anchor words matches in the ASR stream may be opposite to the initial order of words in the CC, as illustrated in Figure 3b. Such temporally conflicting anchor words are discarded (Figure 3b).

The identified anchor words (AW) in the closed caption will then inherit the timestamps of the corresponding tokens in the ASR transcripts. From now on, each AW will be characterized by its final starting and end timestamps, respectively denoted by ST_{AW_F} and ET_{AW_F}.

3) PHRASE ALIGNMENT

For a phrase containing one or multiple AWs we can determine its associated timestamps using the following procedure. We start iterating through its elements in order to determine the first (AW_{FIRST}) and the last (AW_{LAST}) anchor words in terms of temporal location (Figure 4). Using the timestamps of the first AW we can determine the delay of the sentence/phrase starting point, while using the end time of the final AW we can estimate the novel end time for the current sentence/phrase as:

$$ST_{P\_F} = ST_{P\_I} + \left(ST_{AW_{FIRST\_F}} - ST_{AW_{FIRST\_I}}\right), \tag{6}$$

$$ET_{P\_F} = ET_{P\_I} + \left(ET_{AW_{LAST\_F}} - ET_{AW_{LAST\_I}}\right), \tag{7}$$

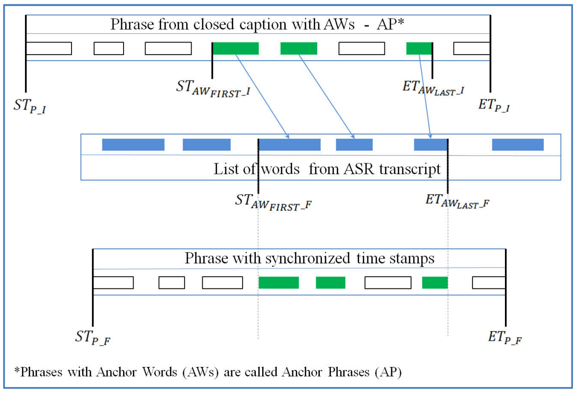
Phrase from closed caption with AWs - AP*
ST_{P_I}
$ST_{AW_{FIRST\_I}}$
$ET_{AW_{LAST\_I}}$
ET_{P_I}
List of words from ASR transcript
$ST_{AW_{FIRST\_F}}$
$ET_{AW_{LAST\_F}}$
Phrase with synchronized time stamps
ST_{P_F}
ET_{P_F}
*Phrases with Anchor Words (AWs) are called Anchor Phrases (AP)

FIGURE 4. Sentence/phrase with anchor words synchronization.

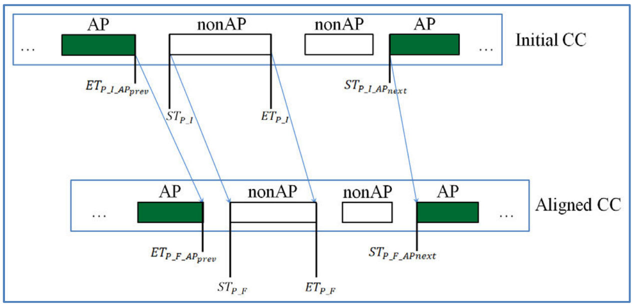

FIGURE 5. Sentence/phrase with no anchor words synchronization.

The aligned sentences/phrases with AWs are denoted as *anchor phrases* (APs).

We need to highlight that: when displaying the closed caption (CC) on the users screen any video player uses only the timestamps assigned to each phrase. So, there is no need to further compute the timestamps associated to each individual word. The sentences/phrases with no anchor words are recursively synchronized based on previously aligned APs and with respect to their original temporal duration in the initial CC (Figure 5).

We considered as reference the $ET_{P\_F\_AP_{prev}}$ of the previous anchor phrase and the $ST_{P\_F\_AP_{next}}$ of the successive anchor phrase.

The novel timestamps for the current phrase are determined as described in the following equations:

$$ST_{P\_F} = ET_{P\_F\_AP_{prev}} + \frac{ST_{P\_F\_APnext} - ET_{P\_F\_AP_{prev}}}{ST_{P\_I\_AP_{next}} - ET_{P\_I\_AP_{prev}}} * \\ * (ST_{P\_I} - ET_{P\_I\_AP_{prev}}), \tag{8}$$

$$ET_{P\_F} = ST_{P\_F\_AP_{next}} + \frac{ST_{P\_F\_APnext} - ET_{P\_F\_AP_{prev}}}{ST_{P\_I\_AP_{next}} - ET_{P\_I\_AP_{prev}}} * \\ * (ET_{P\_I} - ST_{P\_I\_AP_{next}}), \tag{9}$$

Let us note that this procedure insures that no overlap appears between different phrases of the text transcript.

In the end, we make the final adjustment of the phrases timestamps with respect to the audio-visual recommendations. From [23] it can be observed, that the recommended minimum durations for lecture (*i.e.*, the number of seconds necessary for a phrase to be displaced on the screen) depends on the number of characters of the phrase.

Our objective is to adjust the phrases having a temporal duration inferior to the display time specified in the CSA recommendation [23]. Therefore, the system expands the display time of a current phrase whose duration is inferior to a recommended minimum duration for reading. The expanding on the *left* side of the phrase is presented further.

We determine the recommended phrase duration ($REC_{P_{crt}}$), computed based on the number of its characters, and the required bonus duration ($BONUS_{TIME}$) that represents the offset needed to expand the current phrase in either left/right directions:

$$BONUS_{TIME} = \frac{REC_{P_{crt}} - len_{P_{crt}}}{2}, \tag{10}$$

where $len_{P_{crt}}$ represents the temporal duration of the current phrase. The space between the current and previous phrases is defined as:

$$Space_{prev} = ST_{P_{crt}} - ET_{P_{prev}}, \tag{11}$$

If $Space_{prev} > BONUS_{TIME}$ there is sufficient room to expand to the left the current phrase and its starting time is set to:

$$ST_{P_{crt}\_F\_REC} = ST_{P_{crt}\_F} - BONUS_{TIME}, \tag{12}$$

If $Space_{prev} < BONUS_{TIME}$ the maximum available dilatation time on the left DT_{LEFT} is being computed as:

$$DT_{LEFT} = min\left(BONUS_{TIME} - Space_{prev}, \frac{\left(len_{P_{prev}} - REC_{P_{prev}}\right)}{2}\right), \tag{13}$$

if $DT_{LEFT} < 0$, $ST_{P_{crt}\_F\_REC} = ST_{P_{crt}\_F} - \frac{Space_{prev}}{2}$.

If $DT_{LEFT} > 0$:

$$ST_{P_{crt}\_F\_REC} = ST_{P_{crt}\_F} - Space_{prev} - DT_{LEFT}, \tag{14}$$

$$ET_{P_{prev}\_F\_REC} = ET_{P_{prev}\_F} - DT_{LEFT}, \tag{15}$$

Finally, the starting time of the current phrase as well as the ending time of the previous phrase are set to the newly computed recommended values: $ST_{P_{crt}\_F} = ST_{P_{crt}\_F\_REC}$ and $ET_{P_{prev}\_F} = ET_{P_{prev}\_F\_REC}$.

The expanding on the *right* side of the phrase is performed as follows. The space between the current and next phrases is defined as:

$$Space_{next} = ST_{P_{next}} - ET_{P_{crt}}, \tag{16}$$

If $Space_{next} > BONUS_{TIME}$ there is sufficient room to expand to the right the current phrase and its ending time is set to:

$$ET_{P_{crt}\_F\_REC} = ET_{P_{crt}\_F} + BONUS_{TIME}, \tag{17}$$

If $Space_{next} < BONUS_{TIME}$ the maximum available dilatation time on the right DT_{RIGHT} is being computed as:

$$DT_{RIGHT} = min\left(BONUS_{TIME} - Space_{next}, \frac{\left(len_{P_{next}} - REC_{P_{next}}\right)}{2}\right), \tag{18}$$

if $DT_{RIGHT} < 0$, $ET_{P_{crt}\_F\_REC} = ET_{P_{crt}\_F} + \frac{Space_{next}}{2}$.

If $DT_{RIGHT} > 0$:

$$ET_{P_{crt}\_F\_REC} = ET_{P_{crt}\_F} + Space_{next} + DT_{RIGHT}, \tag{19}$$

$$ST_{P_{next}\_F\_REC} = ST_{P_{next}\_F} + DT_{RIGHT}, \tag{20}$$

Finally, the starting time of the current phrase as well as the ending time of the previous phrase are set to the newly computed recommended values: $ST_{P_{next}\_F} = ST_{P_{next}\_F\_REC}$ and $ET_{P_{crt}\_F} = ET_{P_{crt}\_F\_REC}$.

B. SUBTITLE/CLOSED CAPTION POSITIONING

In this section we focus our attention on the subtitle/closed caption positioning on the screen. First, we have specified a set of rules [28] that need to be respected when displaying the graphical text:

(1). The line length is limited to 37 fixed-width (monospaced) characters.

(2). Each line should have a single complete sentence.

(3). The maximum number of lines to be displayed is limited to two.

(4). The subtitles are broken at logical points.

The ideal line-break is a piece of punctuation like a full stop, comma or dash.

However, different from existing video players (*e.g.*, VLC, BS Player, Windows Media Player), which display the subtitle/closed caption on a fixed position (bottom of the screen) regardless of the video content, we perform a complex analysis of the image sequence. Then, the textual information is adjusted in terms of: position, number of lines or line length by taking into account the semantic information existent in the frame.

We start by performing an efficient and accurate text detection using the EAST algorithm [29]. The key element of the system is the neural network model which is trained to directly predict the existence of text instances and

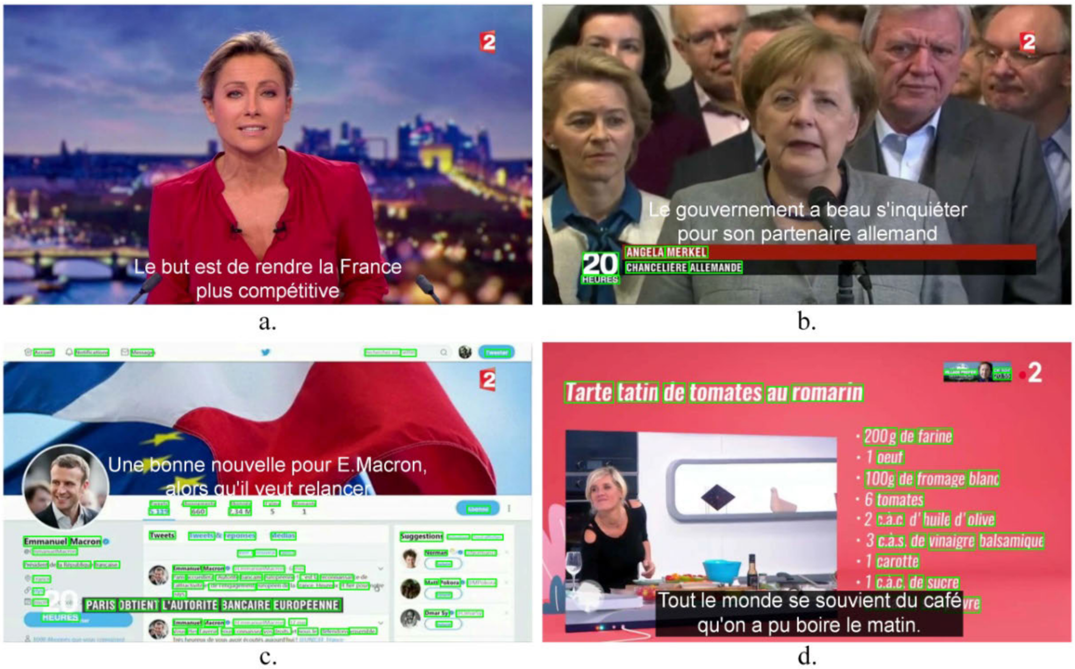
2
Le but est de rendre la France
plus compétitive
a.
2
Le gouvernement a beau s'inquiéter
pour son partenaire allemand
ANGELA MERKEL
CHANCELIÈRE ALLEMANDE
20 HEURES
b.
2
Une bonne nouvelle pour E.Macron,
alors qu'il veut relancer
Emmanuel Macron
Tweets
Suggestions
PARIS OBTIENT L'AUTORITE BANCAIRE EUROPEENNE
c.
2
Tarte tatin de tomates au romarin
• 200g de farine
• 1 oeuf
• 100g de fromage blanc
• 6 tomates
• 2 c.à.c. d'huile d'olive
• 3 c.à.s. de vinaigre balsamique
• 1 carotte
• 1 c.à.c. de sucre
Tout le monde se souvient du café
qu'on a pu boire le matin.
d.

FIGURE 6. Subtitle positioning in various conditions: (a). When no other textual information is present on the video frame; (b). When the scene text is detected on the bottom part of the screen; (c). Scene text is detected on multiple parts of the video frame; (d). When there is no available location on the video frame not overlapping the scene text.

their geometries from full images. The model is a fully-convolutional neural network adapted for text detection that outputs dense per-pixel predictions of words or text lines. In the context of our application, the *scene text* presents additional complexities compared with regular text, such as: multi-orientation or multilingual issues. In the context of the proposed framework we extended the EAST algorithm to work on video streams (instead of static frames). In addition, in order to increase the system's robustness, relevant text regions are tracked between successive frames using the ATLAS approach [30].

Because our major goal is to position the CC within the video stream we decided to discard all detected texts that are not in a horizontal position because this information is not relevant, most of the time, from the perspective of an end user. After the scene text is extracted from the video stream a further analysis is performed in order to determine an appropriate position for the subtitle.

The following cases have been identified and addressed:

1. *No scene text* - when no horizontal text has been detected within the video frame, the closed caption is positioned on the default location (*i.e.*, on the bottom of the viewer screen) as illustrated in Figure 6a.

2. *Closed caption overlapping scene text* - when horizontal text is detected on the bottom part of the video frame (on the default CC location), the closed caption is moved up in the video frame in order not to cover the detected scene text. The novel location is established so that to minimize the distance between the current location and the default position as presented in Figure 6b.

3. *Scene text identified in various parts of the video frame and overlapping with the closed caption* - when horizontal text is detected on the bottom part of the video frame (on the default CC location) and in multiple locations on the frame, the closed caption is moved up on the first available location with the constraint not to overlap any of the detected text (Figure 6c).

4. *Scene text identified in various parts of the video frame and overlapping with the closed caption* - when horizontal text is detected in multiple locations on the video frame, and there is no available position to display subtitle, then the text segment is displayed on the default location but on a black background (Figure 6d).

5. *Scene text overlapping for limited time* - correspond to the case when a scene text is displayed only for a short duration of time (*e.g.*, several seconds), while presenting an overlap with the default CC position. In this situation, the closed caption is placed into the final location that insures no overlap with the scene text (Figure 7b). Even if for some time the CC is above the default position and no overlap with the scene text can be detected (Figure 7a), such a strategy avoids abrupt, short changes in the CC position, which are visually uncomfortable.

IV. EXPERIMENTAL EVALUATION

In order to validate the usefulness of our synchronization and positioning framework in the context of helping the deaf, hearing impaired or hard of hearing assistance, we have set up an evaluation protocol detailed in this section. We notably study the overall system performance with respect to the video genre, number of anchor words or percentage of corrupted words.

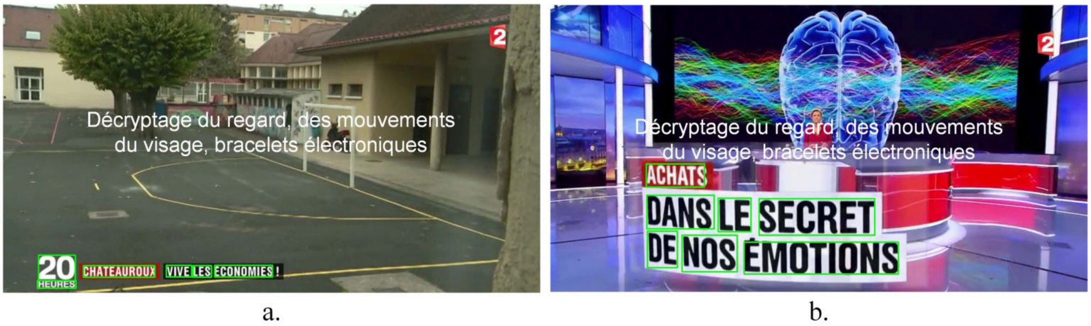
Décryptage du regard, des mouvements
du visage, bracelets électroniques
20 HEURES
CHATEAUROUX
VIVE LES ECONOMIES !
2
a.
Décryptage du regard, des mouvements
du visage, bracelets électroniques
ACHATS
DANS LE SECRET
DE NOS ÉMOTIONS
2
b.

FIGURE 7. Closed caption positioning when a phrase is displayed for several seconds: (a). CC movement with no scene text; (b). Avoid overlapping with scene text.

In addition, we compare our system with a salient state of the art technique [22] and show that the proposed methodology allows significant reduction of the error rate (with more than 55%). Finally, the subjective evaluation performed in real-life scenarios, with a panel of 17 deaf and hearing impaired users is presented and discussed.

A. OBJECTIVE SYSTEM EVALUATION

In order to demonstrate the efficiency of the proposed framework, we have conducted an extensive experimental evaluation, carried out on 30 video sequences, with the duration varying between 20 minutes and 2 hours, recorded at a resolution of 1024 × 576 pixels, at a frame rate of 25 fps. We have considered for evaluation a variety of challenging videos including: TV news, documentaries, movies (including soap opera and sitcoms), TV games and talk shows. We focused our attention on movies with subtitle/ closed caption for which the ground truth timestamps are available in advance. The videos have been selected from the France Télévisions broadcasted documents each category containing six multimedia files.

In order to evaluate objectively the performances of the proposed approach, we have considered the traditional accuracy score, widely used in the state of the art and defined as [31]:

$$Accuracy = \frac{TP}{TP + FP}, \tag{21}$$

where FP denotes the number of false positives (*i.e.,* subtitle segments with incorrect timestamps) and TP represents the number of true positive instances (*i.e.,* subtitle segments that are correctly synchronized).

A subtitle segment is considered as correctly aligned if the following condition is fulfilled:

$$TP = \sum_{i=1}^{N} 1 \left(\max \left(\left| ST_{P\_F_i} - ST_{GT_i} \right|, \right. \right. \\ \left. \left. \times \left| ET_{P\_F_i} - ET_{GT_i} \right| \right) < Th \right) \tag{22}$$

where $1(.)$ is the indicator function that returns one if the condition inside is true or zero otherwise, ST_{GT} and ET_{GT} represent the phrase ground truth start and end timestamps, respectively, N is the total number of phrases existent within the considered subtitle files and Th is a threshold parameter denoting the specified tolerance.

The considered subtitle/closed caption documents have been artificially unsynchronized by randomly modifying the timestamps associated to each text segment based on the following protocol:

(1). A global delay, randomly selected between 10 and 30 seconds, has been inserted at the level of the entire subtitle document.

(2). The timestamps of each textual segment have been altered with an offset between [−15, 15] seconds.

(3). The display time of a phrase has been randomly increased/decreased with a rate between [0.8, 1.2].

(4). No temporal overlap between consecutive captioning segments is allowed. This condition is naturally fulfilled for subtitle documents generated by human transcribers.

The alignment/synchronization results, for various tolerance levels, are presented in Figure 8a. A subtitle segment is considered as being aligned if its timestamps (ST_{P_F} and ET_{P_F}) differ from the corresponding ground truth start and end times with less than the specified tolerance. Also note that the proposed algorithm may offer the possibility of post-processing the resulting alignments to be consistent to the CSA recommendation (*cf.* Section III.A.3). We need to highlight that the ground truth timestamps established by human observers have not considered the audio-visual recommendations.

For clarity in comparisons, the experimental results presented in Figure 8a have not included the final timestamps adjustment with respect to CSA [18]. As it can be observed, for a tolerance of 300 ms the percentage of phrases correctly aligned (Figure 8a) is superior to 93%. Given that a subtitle segment has a minimum duration of display of at least two seconds [18], the choice of this tolerance level is reasonable and may not affect the user viewing experience.

Next, we focused our attention in determining the alignment performances for each category of multimedia documents considered: TV news, documentaries, movies, TV games and talk shows. In Figure 8b we present the experimental results concerning the average percentage of phrases correctly synchronized, per class of documents, when imposing a tolerance of 300ms.

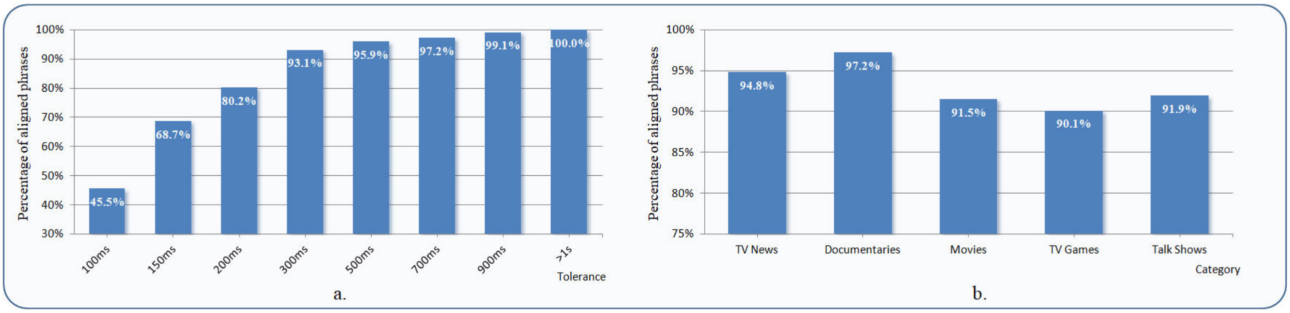

FIGURE 8. Experimental evaluation: (a). Percentage of correctly aligned phrases using the proposed synchronization algorithm; (b). Percentage of correctly aligned phrases per category of multimedia documents when considering a tolerance of 300ms.

TABLE 1. Experimental results of the proposed framework.

Multimedia category	Accuracy (%)	Synchronization error (ms)	Standard deviation (ms)
TV news	94.8	175	135
Documentaries	97.2	159	124
Movies	91.5	208	180
TV games	90.1	227	196
Talk shows	91.9	202	174
TOTAL	**93.1**	**194**	**162**

When analyzing the experimental results, it can be observed that the proposed framework returns the best performances for documentaries with more than 97%, while the lowest accuracy score is obtained for TV games (around 90%). This behavior can be explained by the fact that for documentaries, the audio signal is typically created in the studio, where the actors read a written text and has a reduced degree of noise. In the case of TV game the corresponding speech segments are acquired under unconstrained conditions, with music or background noise which translate to a higher level of imperfect transcripts generated by the ASR.

In order to offer a global performance evaluation of the proposed system, we present in Table 1, for each category of video documents considered, some statistical information, in terms of accuracy, subtitle synchronization error (SE) rate and standard deviation. The SE is defined as:

$$SE = \frac{1}{2N}\sum_{i=1}^{N} \left|ST_{P\_F_i} - ST_{GT_i}\right| + \left|ET_{P\_F_i} - ET_{GT_i}\right|, \tag{23}$$

As it can be observed the experimental results are statistically consistent with an average synchronization error of 194 ms and a standard deviation of 162 ms.

Next, we determined the robustness of our framework in noisy multimedia files. In this context, we have corrupted the unsynchronized transcript by randomly inserting, deleting, and substituting words. The inserted errors are equally distributed within the subtitle/closed caption textual document. The corresponding phrase alignment results are shown in Figure 9a (for a tolerance of 300 ms). As it can be observed even when 10% of the transcribed words are corrupted, the proposed framework is robust enough to provide accurate closed caption alignment. Finally, we have evaluated the impact the number of anchor words has over the system's performance (Figure 9b). In this case, if the ASR result is adequate enough to provide a set of suitable points for the subsequent stages of the synchronization, only minor 1% degradation in performance can be observed when reducing the number of anchor words with 5%. We need to highlight that the original word error rate (WER) achieved by Sphinx, for the French language model, is of 35%.

For a complete evaluation of the proposed framework we have considered also three videos captured from the France Télévisions stations, with both audio-video channels and live generated, unsynchronized, subtitles documents. A manual alignment of the subtitles has been carried out on the three elements considered, thereby establishing the reference timestamps for each textual segment. Figure 10 illustrates the percentage of subtitle segments (in y-axis) with the associated delay in milliseconds (in x-axis) for the original subtitles and for the automatically synchronized versions.

As it can be observed in the original case the subtitle documents have an average delay of 6.93 seconds, while for the synchronized subtitles the delay is reduced to 0.167 seconds.

Finally, we have compared the proposed subtitle synchronization framework with the method introduced in [32]. We have considered the results reported by authors. In [32] the average synchronization delay is 0.303 seconds with a standard deviation of 0.988 seconds. As it can be observed, our approach reduces the average delay with more than 55%.

The subjective evaluation of the approach is presented in the following section.

B. SUBJECTIVE SYSTEM EVALUATION

In order to provide a qualitative system evaluation we have conducted a comprehensive usability study, aiming at comparing the following three paradigms:

(1). Unsynchronized static subtitle (USS) - in this case, the subtitle is unsynchronized and is positioned on a fixed location (at the bottom of the screen);

(2). Synchronized static subtitle (SSS) - the subtitle is synchronized with the audio stream (*cf.* Section III.A.1) and is positioned on the bottom part of the screen;

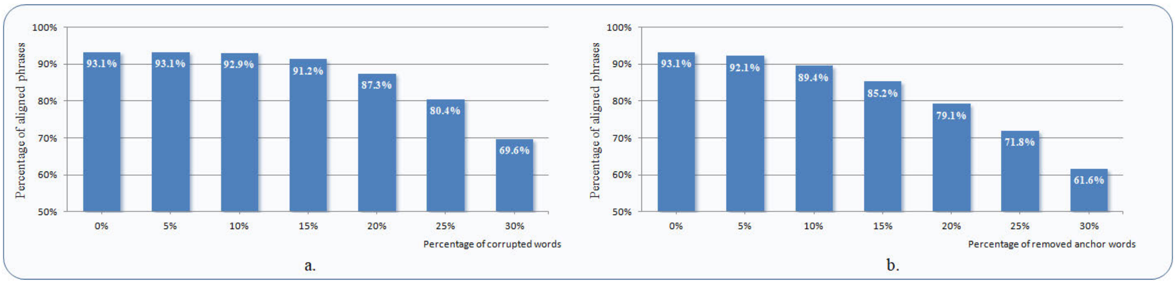

FIGURE 9. System performance variation when: (a). Corrupting the subtitle documents; (b). Modifying the number of anchor words.

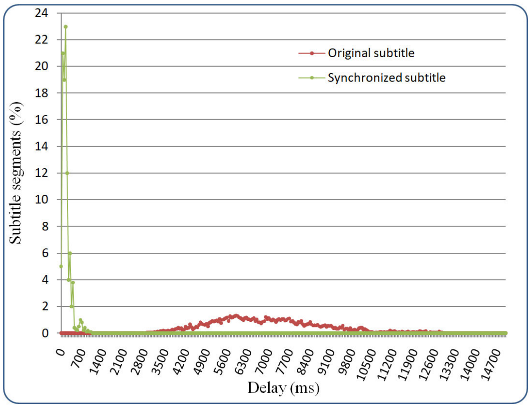

FIGURE 10. Distribution of the delays for the original / automatically synchronized subtitle documents.

(3). Synchronized dynamic subtitle (SDS) - the subtitle is synchronized with respect to the CSA recommendation and its location is dynamically adjusted in order not to occlude important visual content (*i.e.,* textual information incrusted in the video signal).

Concerning the end users, we have recruited 17 anonymous deaf and hearing impaired users. The youngest participant was 18 years old while the oldest had 67 years. All participants have been asked to score (on a scale from A to C – with A indicated highest satisfaction) each of the three versions of multimedia documents based on the overall viewing experience and the eyestrain requirements necessary to follow the subtitle together with the main action. After viewing the videos, the users were asked to respond to the following questions:

(Q1). *Offset*: Is the subtitle display time offset acceptable?

(Q2). *Consistency*: Is there a consistency between the subtitle and the video frames?

(Q3). *Location*: Is the subtitle displayed on the appropriate location?

(Q4). *Display time:* Is there enough time to read the subtitle segments?

The experimental results are presented in Figure 11.

After analyzing the results the following conclusions can be highlighted:

(1). The users definitely preferred the synchronized subtitle strategies over the USS. The users reported that either SSS or SDS are natural to follow and expressed an increased accessibility to multimedia documents with synchronized subtitles/closed captions especially for dynamic video scenes or talk shows.

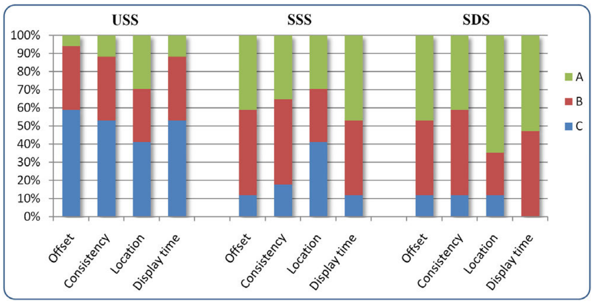

FIGURE 11. Subjective system evaluation with deaf and hearing impaired user.

(2). The use of dynamic subtitles increases the HIP level of understanding over the multimedia documents. In this case the semantic content of the video document is taken into consideration when establishing the location of each textual segment. More specifically, the subtitle does not occlude important visual content incrusted in the video signal.

(3). When adjusting the subtitle timestamps with respect to the CSA recommendation, most HIP indicated that even for highly dynamic video content, there is enough time to both read the subtitle and understand the visual information.

(4). In terms of scores obtained after the subjective systems evaluation of the three methods analyzed we have observed that USS returns the majority of C scores (indicating the lowest satisfaction), the SSS has the majority responses B, while for the SDS the evaluation score has a majority of A.

V. CONCLUSION AND PERSPECTIVES

In this paper we have introduced a novel automatic subtitle synchronization able to automatically align the subtitle/closed caption with the audio stream without requiring *any human* intervention. Moreover, the subtitle text documents and the audio stream are synchronized in a way that fully respects the audiovisual recommendations imposed in France by the *CSA* [18]. In addition, the proposed system performs a high level understanding of the video semantic content in order to position the subtitle within the video frames so that no overlap with other scene textual information occurs. To the best of our knowledge, the proposed system is the first solving this issue in an automatic manner.

The experimental evaluation performed on a large dataset of 30 videos taken from the French national television validates the approach, which is able to return an accuracy score superior to 90%, regardless on the video genre. When compared with other state of the art system [32] our framework reduces the average delay with more than 130ms.

As future work, we envisage to further extend the proposed framework with a multimodal dynamic subtitle positioning system able to detect and recognize the identity of the active speaker and position the subtitle segments in their vicinity. In addition, we plan to conduct a more comprehensive user study evaluation on a larger dataset with hearing-impaired users.

REFERENCES

[1] B. Safadi, M. Sahuguet, and B. Huet, "When textual and visual information join forces for multimedia retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Apr. 2014, pp. 265–272.

[2] *Deafness and Hearing Loss From World Health Organization (WHO)*. Accessed: May 27, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-andhearing-loss

[3] J. E. Garcia, A. Ortega, E. Lleida, T. Lozano, E. Bernues, and D. Sanchez, "Audio and text synchronization for TV news subtitling based on automatic speech recognition," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast.*, May 2009, pp. 1–6.

[4] P. J. Moreno, C. Joerg, J.-M. Van Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *Proc. Int. Conf. Spoken Lang. Process*, Dec. 1998, pp. 2711–2714.

[5] A. Katsamanis, M. P. Black, P. G. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Proc. Workshop New Tools Methods Very-Large Scale Phonetics Res.*, Philadelphia, PA, USA, Jan. 2011, pp. 1–4.

[6] T. J. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," in *Proc. Interspeech*, Sep. 2006, pp. 1–4.

[7] N. Braunschweiler, M. J. F. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. Interspeech*, Sep. 2010, pp. 2222–2225.

[8] A. Haubold and J. R. Kender, "Alignment of speech to highly imperfect text transcriptions," in *Proc. IEEE Multimedia Expo Int. Conf.*, Jul. 2007, pp. 224–227.

[9] S. Hoffmann and B. Pfister, "Text-to-speech alignment of long recordings using universal phone models," in *Proc. Interspeech*, Aug. 2013, pp. 1520–1524.

[10] I. Ahmed and S. K. Kopparapu, "Technique for automatic sentence level alignment of long speech and transcripts," in *Proc. Interspeech*, Aug. 2013, pp. 1516–1519.

[11] M. H. Davel, C. V. Heerden, N. Kleynhans, and E. Barnard, "Efficient harvesting of Internet audio for resource-scarce ASR," in *Proc. Interspeech*, Aug. 2011, pp. 3154–3157.

[12] N. T. Vu, F. Kraus, and T. Schultz, "Rapid building of an ASR system for under-resourced languages based on multilingual unsupervised training," in *Proc. Interspeech*, Aug. 2011, pp. 1–4.

[13] X. Anguera, J. Luque, and C. Gracia, "Audio-to-text alignment for speech recognition with very limited resources," in *Proc. Interspeech*, Sep. 2014, pp. 1405–1409.

[14] *Kaldi a Toolkit for Speech Recognition*. Accessed: Apr. 20, 2021. [Online]. Available: http://kaldi-asr.org/doc/

[15] M. J. F. Gales, K. M. Knill, and A. Ragni, "Unicode-based graphemic systems for limited resource languages," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5186–5190.

[16] G. Bordel, M. Penagarikano, L. J. Rodríguez-Fuentes, A. Álvarez, and A. Varona, "Probabilistic kernels for improved text-to-speech alignment in long audio tracks," *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 126–129, Jan. 2016.

[17] B. Axtell, C. Munteanu, C. Demmans Epp, Y. Aly, and F. Rudzicz, "Touch-supported voice recording to facilitate forced alignment of text and speech in an E-Reading interface," in *Proc. 23rd Int. Conf. Intell. User Interface*, Mar. 2018, pp. 129–140.

[18] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. London, U.K.: Springer, 2014.

[19] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012, doi: 10.1109/MSP.2012.2205597.

[20] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012, doi: 10.1109/TASL.2011.2134090.

[21] P. Żelasko, P. Szymański, J. Mizgajski, A. Szymczak, Y. Carmiel, and N. Dehak, "Punctuation prediction model for conversational speech," in *Proc. Interspeech*, Sep. 2018, pp. 2633–2637.

[22] I. Gonzalez-Carrasco, L. Puente, B. Ruiz-Mezcua, and J. L. Lopez-Cuadrado, "Sub-sync: Automatic synchronization of subtitles in the broadcasting of true live programs in Spanish," *IEEE Access*, vol. 7, pp. 60968–60983, 2019, doi: 10.1109/ACCESS.2019.2915581.

[23] *The Recommendation of the Conseil supérieur de l'audiovisuel (CSA)*. Accessed: Apr. 6, 2021. [Online]. Available: https://www.csa.fr/Arbitrer/Espace-juridique/Les-relations-du-CSA-avec-les-editeurs/Chartes/Charte-relative-a-la-qualite-du-sous-titrage-a-destination-des-personnes-sourdes-ou-malentendantes-Decembre-2011

[24] A. F. Martone, C. M. Taskiran, and E. J. Delp, "Automated closed-captioning using text alignment," *Proc. SPIE*, vol. 5307, pp. 108–116, Dec. 2003.

[25] *The Open Source Speech Recognition Toolkit (CMUSphinx) Project*. Accessed: Apr. 20, 2021. [Online]. Available: https://cmusphinx.github.io/wiki/download/

[26] C. Huang, W. Hsu, and S. Chang, "Automatic closed caption alignment based on speech recognition transcripts," Columbia Advent, Paris, France, Tech. Rep. 005, 2003.

[27] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. Laurent Yoshua Bengio, and A. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," 2017, *arXiv:1701.02720*. [Online]. Available: http://arxiv.org/abs/1701.02720

[28] *BBC Subtitle Guidelines*. Accessed: Apr. 20, 2021. [Online]. Available: http://bbc.github.io/subtitle-guidelines/

[29] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2642–2651.

[30] B. Mocanu, R. Tapu, and T. Zaharia, "Single object tracking using offline trained deep regression networks," in *Proc. 7th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Nov. 2017, pp. 1–6, doi: 10.1109/IPTA.2017.8310091.

[31] B. Mocanu, R. Tapu, and T. Zaharia, "DEEP-SEE FACE: A mobile face recognition system dedicated to visually impaired people," *IEEE Access*, vol. 6, pp. 51975–51985, 2018.

BOGDAN MOCANU (Member, IEEE) received the B.S. degree in electronics, telecommunications and information technology and the Ph.D. degree in electronics and telecommunication from the University Politehnica of Bucharest, Romania, in 2008 and 2011, respectively, and the Ph.D. degree in informatics from University Paris VI-Pierre et Marie Currie, Paris, France, in 2012. Since 2012, he has been a Researcher with the ARTEMIS Department, Institut Mines-Telecom/Telecom Sudparis, France. His major research interests include computer application technology, such as 3D model compression and algorithm analysis in image processing.

RUXANDRA TAPU (Member, IEEE) received the B.S. degree (Hons.) in electronics, telecommunications and information technology and the Ph.D. degree in electronics and telecommunication from the University Politehnica of Bucharest, Romania, in 2008 and 2011, respectively, and the Ph.D. degree (Hons.) in informatics from University Paris VI-Pierre et Marie Currie Paris, France, in 2012. Since 2012, she has been a Senior Researcher with the ARTEMIS Department, Institut Mines-Telecom/Telecom Sudparis, France. Her research interests include content-based video indexing and retrieval, pattern recognition, and machine learning techniques.