ISSN (Online) 2581-9429

IJARSCT

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 6.252

Volume 2, Issue 1, November 2022

A Survey Study on Automatic Subtitle Synchronization and Positioning System for Deaf and Hearing Impaired People

Santosh S Kale[1], Shruti Dhanak[2], Paras Chavan[3], Jay Kakade[4], Prasad Humbe[5]

Guide, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2,3,4,5]
NBN Sinhgad School of Engineering, Pune, Maharashtra, India

Abstract: *In this study, we provide a subtitle synchronisation and placement system intended to improve deaf and hearing-impaired individuals' access to multimedia content. The paper's main contributions are a novel synchronisation algorithm that can reliably align the closed caption with the audio transcript without any human involvement and a timestamp refinement technique that can modify the duration of the subtitle segments in accordance with audiovisual recommendations. Regardless of the kind of video, the experimental evaluation of the strategy on a sizable dataset of 30 films pulled from the French national television verifies the method with average accuracy scores above 90%. The success of our strategy is demonstrated by the subjective assessment of the suggested subtitle synchronization and location system, carried out with real hearing challenged persons.*

Keywords: Automatic Subtitle Synchronization.

I. INTRODUCTION

The number of live programmes that are available and are broadcast online has increased exponentially in recent years. Most European TV broadcasters transmit and share textual information together with the audio and visual signals to make it easier for people to acquire information. Such information is often displayed as closed captions or video subtitles. The linked text document (created by a human transcriber) captures all the semantic value according to the available visual space and slot of time, even when the subtitle/closed caption does not perfectly match the audio (voice) transcription and seldom correlates to the exact word utterance.

The subtitles are offered to satisfy the demands of a sizable population of deaf/hearing impaired persons as well as to translate the conversation in other languages. According to recent figures released by the World Health Organization, hearing impairments are getting more and more widespread in persons over 50, and by 2050, more than 900 million people will have hearing loss [2]. For these persons, a TV program's audio-video material can be accessed most effectively through its subtitles. Due to this, the majority of nations' TV rules, which are enforced globally, require TV content providers to provide closed captions or subtitles for practically all televised shows.

When dealing with live circumstances, the issue becomes urgent.

The audio transcript is created on-the-fly by a regular person in the studio's subtitling environment.

The corresponding written material appears after the spoken snippet by a few seconds. Due to this, most live TV broadcasts have a large lag (up to 30 seconds) between the audio feed and the time the relevant subtitle is prepared to be presented on the screen. As a result, the user is provided with an out of sync closed caption or subtitle. Additionally, depending on where in the subtitle the textual and visual information is presented, there may be a varied delay. The user's ability to comprehend this consequence is seriously hampered.

Program replay is the sole way to get around this restriction, and TV broadcasters use it the most frequently. In this instance, a manual temporal re-alignment procedure is carried out by human experts because the subtitle is already known. However, such a solution requires a lot of time and money.

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:

DOI: 10.48175/IJARSCT-7393

(1). a reliable, automated process for matching the audio channel of a video file with the closed captions and subtitles. The closed caption data generated by human transcribers is sent into the system. Due to the manual manufacturing procedure used to create this data, it has significant misalignments with the audio channel but still provides relevant information. We are able to create a rough transcription of the parsed audio channel using freely accessible automated speech recognition (ASR) software. Because of the ASR transcript's very low level of accuracy, it is insufficient. ASR system performance is also heavily influenced by the language model. They do have the benefit of being completely coordinated with the material, though. The closed caption can be robustly aligned with the audio transcript without any human involvement thanks to the text alignment algorithm that is presented in this paper, synchronizing the caption with the content.

(2) A method of improvement that modifies the lengths of the subtitles to conform to the audiovisual guidelines established in France by the CSA (Conseil Supérieur de l'Audio-visuel).

(3). In addition, to the best of our knowledge, we propose the first framework that carries out a high degree of comprehension of the video content in order to position the subtitle inside the video frames so that they do not encroach on other textual information that could be present in the scene.

The rest of this paper is 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

In order to change the timestamps of the video subtitle, state-of-the-art methods for addressing the issue of audio/subtitle alignment are mostly based on automated speech recognition (ASR) systems that have been previously trained on various datasets and language models. The lack of manually transcribed and aligned training data is the biggest challenge when developing a speech recognizer in a new target language (such as French, Spanish, or Portuguese).

In order to change the timestamps of the video subtitle, state-of-the-art methods for addressing the issue of audio/subtitle alignment are mostly based on automated speech recognition (ASR) systems that have been previously trained on various datasets and language models. The lack of manually transcribed and aligned training data is the biggest challenge when developing a speech recognizer in a new target language (such as French, Spanish, or Portuguese).

The topic of matching extremely lengthy audio recordings to their corresponding written transcripts is initially discussed in [4], one of the earliest studies to address the subject. The core idea behind the suggested technique is based on a voice recognition method that gradually reduces the vocabulary and language model. The authors also provide the notion of islands of confidence, which are likely to be accurately aligned based on an acoustic confidence measure. The alignment between anchors is subsequently refined or optimized through an iterative process.

Strong speech to text synchronisation is the goal of the open source software toolkit known as SailAlign [5]. First, the audio stream is divided up into manageable segments. To evade using a vocal activity detection module, a word is divided into two pieces. After that, a speech recognition algorithm is used to determine the lexical content of each distinct speech segment. Applying what is known as the "minimum number of words criteria," the parts of the audio that may be reliably aligned are determined, with the remaining audio being regarded as unaligned.

Different approaches in the way that those made acquainted in [6], [7] start by preparation a biased expression model (LM) on textual facts in addition to ASR. The developing transcript is joined to the citation paragraph by utilizing active programming. The works named in [8]–[10] show that few in a way theme adjustment is possible when very little news about the prose is accessible. Thus, in [8], erroneous transcripts generated from lectures and performance videos are joined to the visual and audio entertainment transmitted via radio waves stream utilizing a limited set of acoustic parts, discovered heuristically as anchor points. In [9], Hidden Markov Models (HMM) are secondhand in consideration of perform a force adjustment at the sentence level by utilizing a subgroup of short revelation prepared with dossier from various sounds. By utilizing a set of statement models computed on phonetic groups, bureaucracy can trustworthy join paragraph to speech while being healthy against the copy mistakes. Finally, in [10], by misusing the sound unit of speech information in talk and passage transcripts the authors present a method for text adjustment at the

DOI: 10.48175/IJARSCT-7393

sentence level. Another kin of approaches is attracted on adept habits of training the talk acknowledgment foundations so that increase the system strength. Such means involve: the compliance of existent ASR arrangements, trained on the alike vocabulary, but accompanying a various language [11], cross-language bootstrapping [12] or preparation beginning audile models from very limited volumes of manually joined dossier [7].

An ASR order prepared from the very beginning using nearly joined quotation transcripts to the visual and audio entertainment transmitted via radio waves records is introduced in [13]. The preparation dossier (for instance, visual and audio entertainment transmitted via radio waves books, law making speeches) have happened gathered from the Internet. First, the phoneme sequences are joined to the normalized content transcripts through vital programming. Then, the agreement between the letters of a writing system and the graphemes is acquired through a mold of approximation. Finally, the visual and audio entertainment transmitted via radio waves dossier is split into short divisions and the ASR is prepared utilizing the Kaldi toolkit [14]. In [15], a system for construction of graphemes for some word inscribed in unicode is projected. The main idea is to extract the attribute for the graphemes inevitably utilizing the lineaments from the unicode figure description.

Utilizing the countenance from the unicode character writing. From the plans bestowed above it may be observed that private of the cases the synchrony of the visual and audio entertainment transmitted via radio waves tracks accompanying the text transcripts is usually resolved by compulsory alignment acted at the pertaining to speech sounds level. However, aforementioned approaches show fast their limitations when dealing with accompanying lengthened visual and audio entertainment transmitted via radio waves tracks or acoustically erroneous text transcripts. In specific cases, a more advanced reasoning is necessary. To this purpose, in [16], the authors propose utilizing probabilistic kernels (correspondence functions) about misbehavior in consideration of better handbook adjustment. Recently, in [17] in order to handle the instability of abecedarian study and correct the performance of textbook adjustment structures, the authors introduce a human in u.s. city approach. They suggest to use a finger document detective in order to implement a analogy (anchor points). Then, the analogy rewritten all the while the collocated, read-loudly gathering utilizing tablet e-reciters. The opening of deep education algorithms in the circumstances of ASR, have improved the talk plan conduct and strength [16], [18]–[21]. While the increase in accuracy is important, the open ocean education still plays a restricted role in the usual pipelines [22]. Nowadays, the challenge search outact the speech to handbook copy when handling strident environmental environments, accompanying upbringing clutter or when diversified actors are expressive in the in sync.

However, as recorded above, the link middle from two points the label and the phrasing is a case of a classic issue: the adjustment of character sequences. Even though the ASR can conceivably return correct synchronism middle from two points visual and audio entertainment transmitted via radio waves and idea, trustworthy ASR for all portions of the program stream is a difficult project principally on account of extreme instability of the uttered surroundings, speakers and talk types. Spoken atmospheres can change from clean backgrounds (such as, workshop records) to cacophonous settings(for instance, rustic records, talk-argument backdrop sounds that are pleasant, harmonized or belongings). The scholar talking concedes possibility to be top-secret into two bigger classes: account (for instance, the speaker on television benefactor)and willing (for example, interviews captured on place where stocks are bought, debates on talk-shows). In addition, the ASR electronics needs to deal with accompanying unrestrained terminology (exceptionally for films or TVprograms). Any structure loyal to strong label/inscription adjustment endure cautiously allow for possibility aforementioned issues. A second challenge that stands when operating the synchronism search out decide what method to request when the content copy does not competition accurately the uttered conversation of the publishing. In most of the cases, when a news inventoried its the transcripts it kills the linguistic mistakes or added types of foreign conversation uttered again all the while the course of the visual and audio entertainment transmitted via radio waves stream (such as, bref, voila, uh, understanding. . .). Even Challenging are the cases when the person's repetitiveness is very extreme and the assistant cannot trail all communicated facts. In this case, the captioner reformulates the plan in middling phrases that are not absolutely compared accompanying the visual and audio entertainment transmitted via radio waves stream. The triennial challenge that needs expected overcome is moment of truth unavoidable to act the adjustment. In addition, the cost of specific orders endure not be overdone for the video guests. In this paper, we present a novel methods for mechanical synchronism of the captions/subtitles accompanying the visual and audio entertainment transmitted via radio waves-television streams. Compared accompanying added approaches of the most advanced level, we do not focus our consideration on the ASR preparation or on the dossier procurement, but on adept habits to use anchor conversation in consideration of position the idea copy in boisterous settings or accompanying diversified talking characters. In contrast accompanying most advanced level foundations, the projected whole salaam the visual and audio entertainment transmitted via radio waves-ocular pieces of advice meant for one CSA (Conseil Supérieur de l'Audiovisuel) (i.e., Rec. ''Charte relative ála qualité du sous-titrage á goal des personnes sourdesou malentendantes'' [23]) concerning the appropriate period of translation when a phrase is presented on the screen.An supplementary feature of the projected arrangement concerns the dimensional locating of the label in the television frames. To high-quality of our information, the projected method is the first individual that certainly regulates the position of the label established the optic content. This create it likely to prevent affecting subtitles on particular extents of interest that can usually contain beneficial television handbook facts.

III. PROPOSED METHOD

In the following division, we will concern two types of text documents that are assumed expected usable for a same TV program [24]: (1) terminated captions (CC) - that are program transcripts, adapted by a human transcriber, that hold occasion law for every group of 3-10 dispute and grant permission or not be processed in original-occasion; (2) audio talk acknowledgment (ASR)outputs – that denote theme produce instinctively by a speech acknowledgment spreadsheet. In this latest case a time rule is befriended for each word. The objective search out as a matter of usual practice join the closed heading accompanying the transcript of the ASR gain. First, we have completed activity an study of both the independent inscription and the mechanical speech acknowledgment outputs. We have noticed that miscellaneous types of errors can happen. For the CC, four important types of mistakes can appear:(1). orthography mistakes: agrees to words that were misplaced, additional or misspelled apiece captioner;(2). interpretation wrongs: the human assistant paraphrases or replaces the oral communication/words accompanying a akin one;(3). group mistakes: show block of dispute or phrases that are skipped in the finished heading paragraph from the program transcript;(4) periodic dispute or lines: pertain words that are categorized doubly or diversified times by accident. In the ASR productivity three types of mistakes have been labeled:(1). word pronounced the same as another replacement errors: show mistakes generated for one ASR plan that substitutes a correct discussion to its letters of a writing system somewhat the complete discussion;(2). misinterpretation wrongs: the ASR replaces a discussion accompanying a different individual; (3). group wrongs: show a set of dispute (block of words)that are misplaced completely from the visual and audio entertainment transmitted via radio waves stream apiece ASR. Dealing with these miscellaneous types of mistakes for ensuringa healthy CC/ASR adjustment is the key challenge that we attempt to answer accompanying the projected arrangement. The proposed CC/ASR adjustment design is bestowed in Figure 1. The structure takes as recommendation a combined use of several media file that contains the visual and audio entertainment transmitted via radio waves and television streams that are assumed expected absolutely coordinated. Each multimedia file has a offset period (ST) and an end occasion (ET), that delimit the chronology for the considered document. To each combined use of several media file, a shut inscription document is too mixed that involves the text copy of the matching visual and audio entertainment transmitted via radio waves facts. The system exploits three extrinsic elements: a language model toolkit, an mechanical talk acknowledgment order and a quotation detection foundation (in our case established convolutional affecting animate nerve organs networks). The talk recognizer adopts the French speech model, takes as recommendation the audio stream of the aim program and create as yield transcriptions and period codes for the recognized words. The projected adjustment invention reassigns a new timestamp each phrase in terminated heading. The text discovery foundation is planned to recognize the position of existent passage within the program stream. This facts create it attainable to regulate the position of the presented subtitle aforementioned that preventing overlaps accompanying the optical content. Concerning the ASR order, temporary we have adopted the open-beginning CMU Sphinx scheme [25],that offers a fair compromise middle from two points bureaucracy performance and necessary thought/computational possessions. Thus, as proved in the approximate judgment of ASR systems bestowed in [26], completed activity on a set of 30 revelation videos of about 30 brief time period time, mandate error rate (WER) completed by Sphinx is of 35%, that is approximately extreme. However, for legal order that are right recognized, the ASR is capable to correctly capture the timestamps of the visual and audio entertainment transmitted via radio waves text copy. In addition, Sphinx determines an before prepared French sound model. We can wish that by adopting a more

DOI: 10.48175/IJARSCT-7393

complex/strong ASR whole, in the way that one of the arising CNN-located approaches [27], the overall depictions to increase.

REFERENCES

[1]. 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.

[2]. 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.

[3]. 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.

[4]. 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.

[5]. 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.

[6]. 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.

[7]. B. Axtell, C. Munteanu, C. Demmans Epp, Y. Aly, and F. Rudzicz, ‘‘Touchsupported 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.

[8]. 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.

[9]. 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.

[10]. 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.

[11]. 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.

[12]. 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.

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

[14]. T. J. Hazen, ‘‘Automatic alignment and error correction of human generated transcripts for long speech recordings,’’ in Proc. Interspeech, Sep. 2006, pp. 1–4.

[15]. G. E. Dahl, D. Yu, L. Deng, and A. Acero, ‘‘Context-dependent pretrained 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.

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

[17]. 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.

[18]. 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

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

[20]. 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