# AN APPLICATION TO PROVIDE TRANSLATED SUBTITLES AND PICTURES FOR YOUTH ENGLISH LEARNERS USING SPEECH-TO-TEXT AND NLP TECHNIQUES

Harry Cao[1], Yu Sun[2] and Ariel Jiang[3]

[1]SuZou Industry Park Foreign Language School, Jiangsu, China
[2]California State Polytechnic University, CA, 91768, USA
[3]Cornell Tech, Pomona, New York, NY, USA

***ABSTRACT***

*Currently, thousands of free K-12 educational videos exist online with the aim of trying to help young students learn outside of the typical scholastic environment. However, most of these videos are in English, so without subtitles it may be difficult for non-native English-speaking students to fully understand them. These students may need to spend time searching for translations and understanding content, which can distract them from grasping the important concepts within the videos. The state-ofthe- art of speech-to-text and NLP techniques might help this group digest the content of instructional videos more effectively. This paper proposes an application that uses speech-to-text, machine translation, and NLP techniques to generate translated subtitles and visual learning aids for viewers of instructional videos. This video application supports more than 20 languages. We applied our application to some popular online educational videos and conducted a qualitative evaluation of its approach and effectiveness. The results demonstrated that the application could successfully translate the English of the videos into the viewers' native language(s), detect keywords, and display relevant images to further facilitate contextual understanding.*

***KEYWORDS***

*Educational videos, mobile applications, language translation applications*.

## 1. INTRODUCTION

Many instructional videos for language learning do not provide an ideal interface for second language learners, and this is particularly true when it comes to younger students. While at school, teachers often have students watch videos. Since these videos are often shown within a short amount of time, second language learners may not be able to learn new vocabulary words or phrases or understand content, placing them at a disadvantage. Now, students who do not understand all the information they are presented in class can watch online learning videos while at home, usually for free, through sources such as Khan Academy [1]. By using sources such as this, students can review instructional content they missed in class. However, since there are often still no translated subtitles available within these videos, second language learners may still have difficulty digesting the material. With our application, second language learner students can better understand these (video) course materials and their content, and even provide feedback to teachers.

David C. Wyld et al. (Eds): CCNET, NLTM, CICS - 2022
pp. 19-27, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120303

With the aid of artificial intelligence (AI), the technologies that turn the spoken word into written text (speech-to-text) are developing at a pace once thought unimaginable. Recently, Google's speech recognition model broke the 95% threshold for human accuracy, meaning that it has officially become better at recognizing spoken words than most humans [2]. Dating back to the 1950s, the first speech recognition systems were used for translating numbers into text [3]. In 1952, Bell Laboratories designed a system called "Audrey," which could recognize some spoken digits (the names of numbers) when using a single voice. Ten years later, "Shorebox" was able to understand 16 words in English. In the 1980s, the number of words that could be understood using AI went from hundreds to thousands. While the accuracy has been slowly improving, little progress was made until Google launched Google Voice Search. Voice Search was an important milestone for Google's ML-driven voice recognition model, which has so far collected over 230 billion actual voice recordings from users. For now, many of these technology components are shared through Google's publicly accessible library, "Speech-to- Text," e.g., Global vocabulary, Noise robustness, and Speech adaptation, all of which help make transcribing the spoken word possible [4]. These technologies are also applied to Google's popular video uploading site, YouTube, where viewers can turn on "CC subtitles" for most videos to get transcriptions displayed at the bottom. This library of technologies, alongside Google's Cloud Translation [5], serve as the foundation for our application's development.

YouTube's CC subtitles depend on Google's latest speech-to-text technologies to auto-generate English subtitles. However, this approach assumes viewers understand English, since it only provides English subtitles or sometimes, though not often, no subtitles at all. YouTube's implementation is also limited by the fact that videos only auto-generate subtitles within the same language used in the videos, with few other languages being supported. Other video transcription sites such as "Rev" [6] and "Scribie" [7] provide a sophisticated level of translation, but require user registration and membership fees that potentially make them less accessible to younger students. One could argue that it is reasonable to charge transcription fees since longer videos would require lots of computational resources [8]. However, most educational videos are short, so it may be beneficial to have a service that offers the transcription of shorter videos (of less than ten minutes) via upload for students only. In addition, other sites do not provide images to help with context. When students need to learn a new word, visual assistance has proven to be an efficient learning tool to understanding and memorization [9]. Science videos frequently use unfamiliar words or technical jargon that students find difficult to visualize. Even if students use a dictionary, this requires additional time, and definitions may still be still be too short or obscure to grasp the required context. If the transcription includes images, however, students don't have to go out of their way to search for meanings and can immediately get a mental picture of the concepts being explained. Image insertion technology is in high demand, and as a result is also very accessible. Currently, there are many video editing sites (e.g., "Kapwing" [10] and "Pixiko" [11]) as well as applications that allow users to insert images at a specified time during viewing. Unlike these sites, our application automatically searches for meaningful words and inserts them into videos to provide visual aids for students, quickly and conveniently.

Our application's main function is to take a short video as input and generate a streamlined final product consisting of three parts: a video transcription, a video translation, and key images for the video. First, we used speech-recognition technology to transcribe existing videos in English. Secondly, we translated the videos into users' native languages and embedded subtitles relevant to these languages. Then the videos were successfully translated into various languages of choice using speech-to-text and machine translation technologies. As previously mentioned, there are several existing services that may also produce video translations, but very few are free and focus exclusively on short, educational videos for students. Finally, we used NLP techniques to analyze the English transcripts and extract keywords and phrases based on a certain set of algorithms that are empirically proven to be the most accurate for this purpose.

After these keywords were found, we inserted related images back into the original videos. This will allow these images to appear briefly whenever the relevant word(s) are spoken to help students clarify meanings and generate mental images of whatever is being discussed. In order to evaluate the performance of our application, we believed it best to conduct experiments arising from two angles: (1) the accuracy of speech-to-text and translation done by our system and (2) user experience. First, we tested the accuracy of speech-to-text for our application. We expected the best metric for this to be numeric accuracy and the gold standard method to be human (speech). For a selection of educational short videos, we let native speakers watch and record what they heard, since humans are very accurate in terms of identifying the words they hear in their own language. We then fed the video into our application and compared our transcription with the human's, word for word, and recorded the accuracy of each. We also conducted similar experiments to test the accuracy of our translation and the combination of both. Since our application is made to serve young English learners, their progress in learning was an important metric of success for us. We invited primary school students who do not speak English well to watch a selection of educational videos in English that they would normally encounter in either a home or school environment. Afterwards, we put the same videos through our translation system and let students re-watch them. We then surveyed the students about their comparative experiences, specifically asking how our embedded images might have helped their understanding of unfamiliar words or contexts.

To outline the rest of this paper, Section 2 provides details on the challenges we met during the experiment and while designing the sample; Section 3 focuses on the details of our solutions corresponding to the challenges mentioned in Section 2; Section 4 presents details relevant to our experiment, followed by related work in Section 5. Finally, Section 6 provides concluding remarks, as well as possible directions for this project in the future.

## 2. CHALLENGES

In order to devise an application that uses speech-to-text, machine translation, and NLP techniques to generate translated subtitles and visual learning aids for viewers of instructional videos, a few challenges were identified as follows.

### 2.1. Challenge 1: Inferring Sentence Breaks within Transcribed Text

The first challenge we faced was how to yield an accurate translation of our English transcripts into our user's native language. Specifically, there is the difficulty of inferring the meaning of a given sentence, since text created from speech-to-text is oftentimes just a continuation of words instead of naturally segmented sentences. Inferring the breakpoint of sentences from a sequence of words without punctuation is a known difficulty surrounding NLP. First, one has to infer the correct number of breakpoints given a sequence of words. Second, one has to find the probability of a breakpoint being inserted between word $w_i$ and word $w_j$ for all $i$ and $j$. There have been few studies done on this topic [12], mainly because it is rare for people to have to break up a sequence of words without punctuation or capitalization.

### 2.2. Challenge 2: Extraction of Keywords

The second challenge we faced was the extraction of keywords from the transcripts. This task had two goals: (1) extracting keywords relevant to the videos, and (2) extracting keywords related to the key concepts of the videos that students are likely encountering for the first time. The first goal can be achieved through established NLP techniques [13], but when combined with the second goal, the problem becomes more difficult to solve. For example, for a video providing a

biological explanation of how a human body works, although the word “human” might appear multiple times and would be a sensible guess as to the topic of the video, key scientific words like “heterozygous” or “oligosaccharide” that only appear once or twice would likely be neglected by traditional NLP methods of choosing keywords or key phrases.

### 2.3. Challenge 3: Selection of Images

The third challenge we faced was the selection of images. Because we intend for our application to insert images into videos as a fully automated process, the selection of images must also be automated. This means that we could only use keywords extracted from the videos themselves to conduct the selection and searching process. This process can be tricky as a simple word like “market” can have different meanings. Even if we know the implied meaning within the context is “economy,” it would still be impossible to identify this meaning through images and pick the correct one. To solve this challenge, more pre-processing needs to be done before the searching phase, e.g., adding the word “grocery” before “market” to emphasize one meaning over another.

## 3. SOLUTION

### 3.1. Overview of the Solution

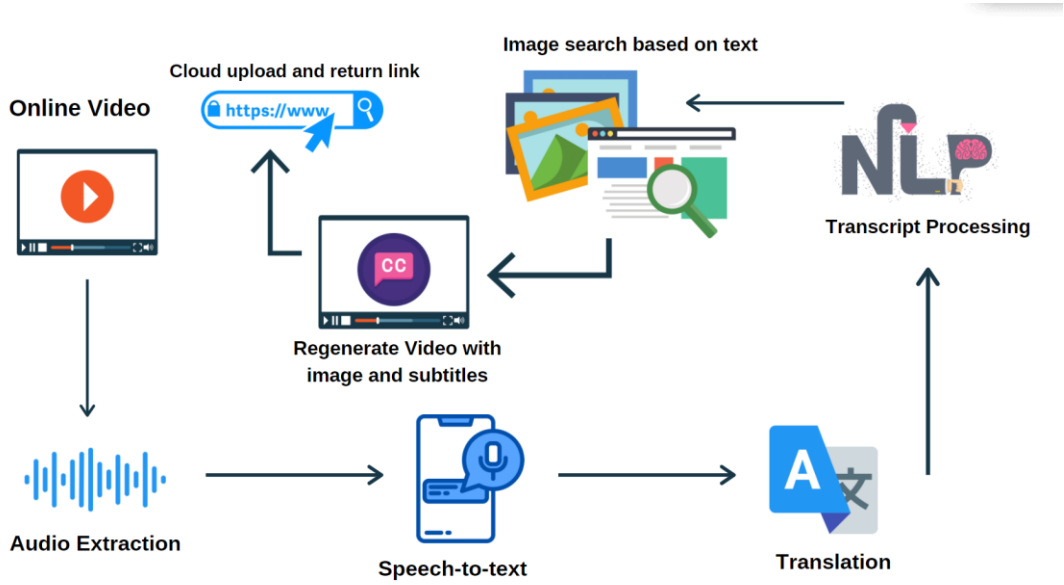


Figure 1. Overview

The application initializes itself and internally creates a Cloud Manager that controls the connection with several third-party APIs (firebase, speech-to-text, translation, image search). After the video is uploaded, it is first processed with audio extraction through efficient video editing tools such as ffmpeg. Then, the audio is uploaded to cloud storage through Cloud Manager to do speech-to-text transcription using Google’s library. During the transcription phase, three types of data are stored: (1) the transcript of the video in English, (2) the “srt” format of the transcript where “srt” is a human-readable file format that stores text sequentially, along with the timing information [14]. In our case, the text we store is the original transcript after getting translated into the user's native language using machine translation provided by a third-party tool. (3) We also store the timeline of the video during the processing phase. The timeline of the video is stored as a dictionary object in our Python implementation, where keys are words spoken

during the video and values are a list of timestamps during which such words appear. This dictionary becomes useful later on, when we need to insert images back into the video at the time(s) they appear. At this point, the transcription of the video is already complete. Next, the system analyzes the English transcripts using NLP technologies and extracts the keywords and phrases that are both representational of the video and add values to the user's knowledge base. We then search the web to find related images for these keywords and insert them back at the time they appear, using the timeline dictionary we generated earlier. After the video's subtitles and images are both embedded back into the original video, we upload the video to the cloud using Cloud Manager and return the cloud link to the user.

## 3.2. Component Breakdown

### 3.2.1. Speech-to-Text Transcription and Translation

These two components are the foundation of our application and are done sequentially in the same iteration using Google's Speech-to-Text and Translate APIs. After the API starts running and listening to the audio, it generates a streamline of results. For every ten words spoken in the stream, we first transcribe the original English words, then translate and record the translation in "srt" format. Meanwhile, the exact timeline of the video is also recorded. The code snippet shows the nested "for" loop where we process each word and store its timestamp to our dictionary for future processing, as mentioned in the previous section. Figure 1 shows where in the sequence the video is regenerated with subtitles and images using tools like ffmpeg to insert "srt."

### 3.2.2. Transcripts Analysis

For transcripts analysis, our goal is to extract the most relevant and valuable words that will help users both learn new concepts and expand their vocabulary or knowledge base. We perform basic text processing techniques such as stopwords removal, tokenization, and tagging sequences with part-of-speech tags all using Python's nltk library. We discovered that words with tags such as "NNP" and "NNPS" (proper nouns) tend to be more valuable and represent the video better, since they are less likely to be generic. Therefore, we applied more significance to these words. Also, we used a sublinear term frequency approach to normalize frequent words, meaning that instead of using the raw count of the terms, we instead used a logarithm of the count plus one. This helped us prevent words that are not stopwords but common within the video, such as "human," from taking too much significance. The number of keywords also depends on the video's length, since we output a number of keywords based on a 1:2 ratio, meaning one minute of a video equals 2 keyword outputs.

Figure 2. Video frame showing an embedded image

### 3.2.3. Image Selection and Insertion

After retrieving the important vocabularies from performing text processing, we use Google's search API to search for the most relevant images and download them. We then use our timeline dictionary to construct an inverted dictionary where each key is the time an image should appear on the screen and the values are the corresponding images. We experimented with various durations of time for images to remain on the screen and settled on three seconds as our final implementation. We also took into consideration the possibility of image overlap, and instituted a minimum one second gap between image displays. After these steps are completed, we insert the images back into the original video using openCV. Depending on the original video frame rate and dimensions, these images will appear with modified heights and widths such that they always only occupy at most . of the screen. After the images and subtitles are both embedded into the original video, we upload the video back to firebase storage and return the publicly accessible link to the user.

## 4. EXPERIMENT

### 4.1. Experiment 1

To evaluate the accuracy of speech-to-text, we used three scientific videos which do not have English subtitles provided and let our application do the transcriptions. Then we invited five teachers to write down the text spoken in the videos. We summarized and unified the data of several teachers, and regarded this as the correct transcript. We then compared this data with the computer speech-to-text transcriptions. We counted all the words that had appeared in the videos, and extracted the words that were different from the teachers' transcriptions and our application had transformed. We then counted the words that were different between the application and the teachers and divided this number by the total number of words to calculate the rate of error. One hundred minus the rate of error number is the accuracy rate of our application. With this margin of error, the application is able to effectively interpret the verbal context of the videos.

Table 1. Deviation rates for accuracy (application versus human)

| Video name | Number of participants who viewed | Rate of deviation between application and participants |
|---|---|---|
| Crash Course | 5 | 3.1% |
| Organic Chemistry Preview | 5 | 1.7% |
| Psychological Research: Crash Course Psychology #2 | 5 | 5.2% |
| Biology: Cell Structure \| Nucleus Medical Media | 5 | 3.9% |

### 4.2. Experiment 2

We showed three children who don't speak English a scientific video containing a lot of jargon and no subtitles. We chose videos that were less than three minutes long and only explained the basic concepts. We allowed the children to take notes while watching the video so that they might memorize the content. After this, we asked the children to give us a brief summary of the video,

which they were unable to do. We then played the same video using our application. After viewing, all three children were able to provide a brief summary of the video. When evaluating these summaries, we used three standards of ability: clarity, understanding of key concepts, and ability to reiterate. The range of marks given were between 0-5. For clarity, 0 meant that the child had no idea of the main topic, while 5 meant that they could explain the main topic well. For understanding of key concepts, 0 meant that the child could not explain the key concepts, while 5 meant that they could explain them well. For ability to reiterate, 0 meant that the child could not explain the video without notes, while 5 meant that they could explain the video well to others.

Table 2. Respective scoring for clarity, understanding of key concepts, and ability to reiterate

| Name of video | Average scores before use application | Summary after use application |
|---|---|---|
| Photosynthesis: Crash Course Biology #8 | ● clarity: 1.2<br>● understanding of key concept: 1.53<br>● ability to reiterate: 0.24 | ● clarity: 4.76<br>● understanding of key concept: 4.87<br>● ability to reiterate: 3.98 |
| Simple Harmonic Motion: Crash Course Physics #16 | ● clarity: 2.87<br>● understanding of key concept: 0.77<br>● ability to reiterate: 0.2 | ● clarity: 4.85<br>● understanding of key concept: 4.67<br>● ability to reiterate: 4.01 |

### 4.3. Analysis

The result of the first experiment shows that there is a high degree of overlap between what people hear and what our application processed, which shows that the speech-to-text feature of our application has an acceptable rate of error. In the second experiment, by contrasting the input from children using the application versus the control group, we can see that our application greatly improved the understanding of the video for non-native English speaking children.

## 5. RELATED WORK

The paper, “Review of Speech-to-Text Recognition Technology for enhancing learning” [15] introduced how speech-to-text technology applies to enhanced learning, especially for nonnative English speakers. Shadiev [16] has shown that non-native English speakers take advantage of nineteen strategies in using STR-generated transcripts. As a result, participants are able to use them to study, compose summaries, and generate their own ideas, which helps them receive higher scores.

The paper, “Review of studies on recognition technologies and their applications used to assist learning in instruction,” [17] reviews and summarizes studies on recognition technologies within the last ten years. Technologies using haptic input were employed mostly within science based education [18]. Students use this technology to take notes during lectures and enhance understanding when information is missed or misunderstood, or for completing homework assignments.

The paper, “Investigating the effectiveness of speech-to-text recognition applications on learning performance, attention, and meditation” [19] examines the effectiveness of speech-totext recognition technology as a learning and concentration aid to encourage a calm state of mind. In this experiment, students who do not have STR (speech-to-text recognition), when unable to

understand their lectures, feel stressed and lose concentration. Meanwhile, students who used STR felt less pressure while trying to learn, since they could properly read and understand texts.

Our goal is to make an application that helps children who don't understand English well to be able to understand and summarize videos presented in English. Speech-to-text technology can help children understand the content of instructional videos in English with less pressure so they can enhance their learning skills.

## 6. CONCLUSION AND FUTURE WORK

In this paper, we proposed an application that automatically transcribes online educational videos and provides contextual visual aids to help K-12 non-native English speakers understand the content better and faster. With the help of speech recognition technology, we built this application's foundation using advanced speech-to-text and machine translation technologies provided by Google. For visual aids, such as images, we effectively applied text processing techniques to first extract key information from the transcript(s), then embedded these images back into the videos to help students obtain a visualization of the video's important concepts. During the implementation phase, we met with a few challenges, including translation inaccuracy and the unpredictability of the image contents as mentioned in Section 2. We were able to eliminate some of the effects these issues had on our application by modifying the system to avoid them. In the end, our evaluation results indicate that the application is effective. In our first experiment, we observed that the error rate of our speech-to-text transcription is negligible and did not seem to affect people's ability to understand the content of the video(s). In our second experiment, we observed that non-English speaking children were better able to summarize videos when using our application versus a control group that viewed the videos without it. This is a good indication that our application does in fact help non-English speaking viewers better understand instructional videos presented in English.

Our application cannot guarantee complete accuracy, however, when generating subtitles for medium to longer-length videos. In addition, many English words have more than one meaning, so there will almost always be errors in translation. While our application also adds images to help understand the most important keywords of the videos, we couldn't predict what these images would be, since the application chooses the first image found by Google and inserts it into the video. While testing the application, some images would appear with strange words or pictures. Although this doesn't affect the aid provided by the images, overall, it doesn't provide a perfect watching experience.

We hope we can further adapt and evolve our application to reduce such errors. We would like our application to be able to differentiate which images are valid and which are not, e.g., which images best illustrate the context and subject matter of the videos. The accuracy of the machine translation could also be improved so that non-native English speakers can understand the videos even more clearly.

## REFERENCES

[1] Free online courses, lessons & practice. (n.d.). Retrieved May 01, 2021, from https://www.khanacademy.org/

[2] A brief history of speech recognition. (n.d.). Retrieved from https://sonix.ai/history-of-speechrecognition

[3] The machines that learned to listen. (n.d.). Retrieved from https://www.bbc.com/future/article/20170214-the-machines-that-learned-to-listen

[4] Speech-to-Text: Automatic Speech Recognition. Google Cloud. (n.d.). Retrieved from https://cloud.google.com/speech-to-text
[5] Cloud Translation. Google Cloud. (n.d.). Retrieved from https://cloud.google.com/translate
[6] Rev Speech-to-Text Services: Convert Audio & Video to Text. (n.d.). Retrieved from http://www.rev.com/
[7] Corporation, C. (n.d.). Transcribe your audio/video files rates starting at $0.10/min. Retrieved from https://scribie.com/
[8] Goddard, W. (2021, March 17). Speech Recognition Algorithm - Brought to you by ITChronicles. Retrieved from https://itchronicles.com/speech-to-text/speech-recognition algorithm/
[9] Vavra, Karen L., et al. "Visualization in science education." *Alberta Science Education Journal* 41.1 (2011): 22-30.
[10] Make Something Awesome. (n.d.). Retrieved from https://www.kapwing.com/
[11] Online video editor: Improve your videos in a few clicks. (n.d.). Retrieved from https://pixiko.com/
[12] Barrows Jr, Randolph C., M. Busuioc, and Carol Friedman. "Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches." *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2000.
[13] Aone, Chinatsu, et al. "A scalable summarization system using robust NLP." *Intelligent Scalable Text Summarization*. 1997.
[14] What is an SRT File? How to Create & Use SRT Files. (2021, March 09). Retrieved from https://www.rev.com/blog/resources/what-is-an-srt-file-format-create-use-srt-files
[15] Rustam Shadiev, et al. “Review of Speech-to-Text Recognition Technology for Enhancing Learning.” *Journal of Educational Technology & Society*, vol. 17, no. 4, 2014, pp. 65–84. *JSTOR*, www.jstor.org/stable/jeductechsoci.17.4.65. Accessed 1 May 2021.
[16] Shadiev, R., Hwang, W. Y., & Huang, Y. M. (in press). Investigating applications of speech to text recognition for face to face seminar to assist learning of non-native English participants. Technology, Pedagogy and Education.
[17] Shadiev, Rustam, et al. “Review of Studies on Recognition Technologies and Their Applications Used to Assist Learning and Instruction.” *Educational Technology & Society*, vol. 23, no. 4, 2020, pp. 59–74. *JSTOR*, www.jstor.org/stable/26981744. Accessed 1 May 2021.
[18] Neri, L., Noguez, J., Robledo-Rella, V., Escobar-Castillejos, D., & Gonzalez-Nucamendi, A. (2018). Teaching of classical mechanics concepts using visuo-haptic simulators. Educational Technology & Society, 21(2), 85–97.
[19] Shadiev, Rustam, et al. “Investigating the Effectiveness of Speech-to-Text Recognition Applications on Learning Performance, Attention, and Meditation.” *Educational Technology Research and Development*, vol. 65, no. 5, 2017, pp. 1239–1261. *JSTOR*, www.jstor.org/stable/45018724. Accessed 1 May 2021.