

A transforming method of video materials into the listener language

Ihor Bandura¹, Oleksandr Osolinskyi¹, Roman Komarnytsky¹

¹ West Ukrainian National University, 11 Lvivska St, Ternopil 46009, Ukraine

Abstract

Currently is a boom in information technology, which cover all industries and spheres of human activity. Online conferences, live video broadcasts, which allow people to exchange thoughts, ideas, knowledge, etc., also play an important role in information technology. The opinions of other people have always been an important part of information. If looks at video blogging, there are many people involved in this field who are engaged in various studies and conduct their own online broadcasts or record videos where they share various useful information that would be interesting to hear from people from different parts of the world. Posting videos in only one language, bloggers lose many viewers who could help them implement some projects, monetize their products, etc.

In this paper proposed a method for converting video into the listener's language using artificial neural networks and describes the entire process from splitting video into sound and video, translating sound into text, and translating text into the listener's language.

Keywords

Transformation, video materials, speech recognition, text synthesizing, translation, neural networks

1. Introduction

During the quarantine, the demand for online conferences, online video viewing, and remote learning using systems such as Zoom has certainly increased, "Google Meet, etc. Most companies have successfully transferred all their employees to remote work. And even after the quarantine ended, many companies decided to continue working remotely and became more loyal to hiring employees from other countries, but this requires that the employee, of course, know the language of the country where the company is located to be able to successfully communicate with colleagues and other employees of that company.

If large companies often refuse to localize their products into multiple languages due to high costs, which will ultimately increase the product costs, then what can we say about people who do their work as a hobby and are not ready to spend a lot of money on translating their own video materials into different languages.

It is clear that machine speech recognition, translation, and voice-over will not replace humans on 100 percent, as the quality will be much lower, but the goal is not to fully automate the translation of video materials, but to make it much easier, simpler, and reduce the time required for translation, and as these issues are resolved, the cost of localization may decrease. Using the method of transforming video materials into the user's language will allow companies to integrate many more languages into their products, which will increase the number of people interested in this product. The same applies not only to entertainment, but also to the translation of video materials for online learning and any video in general.


The First International Workshop of Young Scientists on Artificial Intelligence for Sustainable Development; May 10-11, 2024, Ternopil, Ukraine

✉ igor.bandura.3@gmail.com (I. Bandura); oso@wunu.edu.ua (O. Osolinskyi); komarnytsky.roman@gmail.com (R. Komarnytsky)

ORCID 0009-0009-6494-1647 (I. Bandura); 0000-0002-0136-395X (O. Osolinskyi); 0009-0008-2794-5107 (R. Komarnytsky)



© 2024 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

For people who blog or record different video formats, it will be much easier to localize them for more people, as the cost of localization with this service can be much lower, or they can try to localize their own products themselves using the method proposed in the qualification paper.

The transformation of video materials into another language takes place in three steps, i.e. in order to translate a video, needs to recognize the language from the video, namely convert it into text, translate the text that was recognized in the previous step, and in the last step, needs to use machine voice over [1] to create an audio file with the translated text and attach it to the video. Automatic speech recognition uses technology to convert speech signals into a sequence of words or other linguistic units using an algorithm implemented in a computer program. Today's speech recognition systems are capable of understanding speech input for dictionaries containing thousands of words in an operational environment. The speech signal conveys two important types of information, namely the content of the speech and the gender of the person whose speech was recognized. Speech recognizers are aimed at extracting lexical information from the speech signal, regardless of the frequencies and the way the speaker speaks.

2. Problem Statement

The transformation of video materials into the user's language on a hyper-converged platform represents a critical challenge in the current landscape of digital media. As content production and consumption become increasingly globalized, the ability to efficiently convert video content into multiple languages becomes indispensable for reaching a broader audience. This challenge is compounded in environments where resources and access to advanced translation services are limited.

Existing methods of video transformation often involve complex, resource-intensive processes that may not be feasible for all users, particularly those with limited technical expertise or in resource-constrained settings. Additionally, the quality of transformation can vary significantly, affecting the accessibility and usability of transformed content. There is a clear need for an innovative approach that simplifies the video transformation process while maintaining high standards of accuracy and user-friendliness.

The proposed method aims to address these challenges by developing a user-friendly, efficient, and scalable solution to transform video materials into the user's language using a hyper-converged platform. This method will leverage advanced computational techniques and user-centered design principles to ensure that the transformed videos are accessible and valuable to a diverse global audience. The successful implementation of this method has the potential to significantly enhance the way educational, informational, and entertainment content is consumed across different linguistic and cultural groups, thereby broadening the impact of digital media globally.

3. Related works

The video transformation of materials into the user's language goes through several stages, the main ones being: speech recognition, text translation, and machine voice over.

Since each stage depends on the previous one, the quality of the output data of each stage must be high. Only in this case will the output result of the transformation be of really high quality and meet the user's requirements.

The authors of [2] note that it is extremely important for machine speech recognition to also recognize the intonation with which a particular phrase is pronounced, because this will make it possible to correctly place punctuation marks in the source text. These punctuation marks must be preserved during text translation and will be needed for further pronunciation of the translated phrase in the user's language.

In the work [3] authors also concluded that in order to correctly determine the emotional color of the phrase and place punctuation marks in the phrase, it is necessary to use the "fundamental tone" parameter in the semantic analysis [4] of speech signals because it is the strongest

harmonic in the spectrum of the speech signal, determined by the frequency of oscillations of the human vocal cords when pronouncing vowels and voiced consonants. For detect whether a sentence is interrogative by checking the frequency of the phrase, for example, if the frequency increases at the end of the phrase, it means that the phrase is interrogative, if the frequency remains low and stable, it means that the phrase was spoken with an affirmative tone, and if the frequency is too high, it means that the phrase is exclamatory and an exclamation mark should be placed at the end of the sentence.

The frequency with which the phrase was spoken can also be used to determine the speaker's gender, since the basic tone of human speech is in the range of approximately 60-150 Hz for men and 150-300 Hz for women.

That's why in the work [5] conducted a study in where obtained a neural network model for gender recognition. This approach will work much better for determining gender in speech recognition because it is based on many different features, which makes the task of determining the speaker's gender much more efficient. This approach will produce correct data much more often because it is based not only on the frequency at which the speaker speaks, but also on many different features.

Since the stage of speech recognition and its transformation into text is the most important stage because it outputs the text with which the system will work in the future and the subsequent stages of audio translation are highly dependent on it.

In [6] a study was conducted that showed that in order to make speech recognition as good as possible, it is necessary to use appropriate trained models for the desired language and subject area. Free acoustic models for speech recognition can be found freely available on the Internet, but these models are not perfect and still need to be trained to meet the needs of the project. Therefore, the best possible models for speech recognition should be used.

Text translation is also very important in the process of transforming audio materials into the user's language, since it is the source text that will be voiced by the machine voice system. However, as noted in [7] there will always be many problems at the translation stage and it will always be of insufficient quality, since machine translation is still not of high quality and very often the source text may contain grammatical errors, linguistic errors, etc.

For text translation, as well as for speech recognition, a neural network model for a translation system is important. It is this model that stores the vocabulary and grammatical rules of the target language, which are used to compose the source text. The model can be trained, expanded, and refined, but this will still not be enough, because in order to translate a literary text or certain artistic techniques and descriptions, you will always need a person who can recognize similar emotional colors.

Early methods of transforming human speech into text focused on manual feature extraction and conventional methods such as mixed Gaussian models, also known as GMM [8], the Dynamic Time Warping algorithm, for which there is also a well-known abbreviation DTW [9], and Hidden Markov Models, commonly referred to as HMM [10]. And more recently, neural networks such as recurrent neural networks, abbreviated as RNN [11], convolutional neural networks, also known as CNN [12], and in recent years, neural networks such as transformers [13], which have been applied in automated speech recognition methods, have achieved great success in them and have shown good performance.

Since the best check of the quality of a translated text is only a human being, the best solution to improve the quality of a text translation system is not only to change the neural network model to a better one, but also to provide the end user with access to view and modify the translated text. In this case, the end user will be able to check the translated text and correct errors if any.

Thus, it is advisable to use the full speech synthesis method to transform video materials into the user's language, since the result after speech recognition and speech translation may be different and it is necessary to adapt to the phrase conveyed by the text translation stage.

While researching analogs of the method of transforming video materials into the user's language, founded several services that allow for such a transformation. But main focus was to service «IconnectFx» [14].

4. A method for transforming video materials into user speech

The method of transforming video materials into the user's language consists of several steps:

- extract the audio file from the uploaded video material;
- recognize speech using neural networks from the audio file, i.e. transform the extracted audio file into text;
- translate the text into another language chosen by the user;
- transform the translated text into an audio file, i.e., voice this text;
- separate human voices from other background sounds using artificial intelligence in the original audio file;
- using software tools that allow you to work with video and audio files, reduce the volume of people's voices in the original audio file;
- combine the audio file of the spoken translated text, namely: an audio file with the volume of the original voices reduced, an original audio file with background sounds, and an original video file without sounds.

The transformation will result in a video file translated into another language.

Since at the last stage, before merging the files, human voices are separated from background sounds using artificial intelligence, all background sounds, such as knocking, music, pen clicks, and similar sounds, will not be lost; the file will not be emotionally colored.

It is also possible to leave the original voices at a lower volume, if the user so desires.

Thus, the main goal of automated speech recognition systems is to transform an input audio signal $x = (x_1, x_2, \dots, x_t)$, with a specified length T , into a sequence of words or symbols $y = (y_1, y_2, \dots, y_n)$, where n is the number of words that the system has recognized from the input audio signal. All characters and words returned by the speech recognition system are a subset of the set of characters and words in the dictionary.

All speech recognition methods have the same working principle, which consists of the following steps

processing and normalization of the input signal, also called pre-processing;

- feature extraction;
- classification;
- Speech modeling.

Since transformer neural networks have achieved significant improvements at the moment, they are of good quality as a result of their use in various speech recognition methods and translation of phrases into other languages. Transformer neural network models designed for speech recognition, and it usually based on an encoder-decoder architecture similar to seq2seq models [15]. If we look at them in more detail, they are based on a self-attention mechanism instead of recurrence, which was used in recurrent neural networks. The self-feedback mechanism is one of many techniques for training neural networks. A self-focused mechanism can pay attention to different positions in a sequence and extract meaningful representations. A self-focused mechanism uses three parameters: queries, values, and keys. Let's denote queries as Q , values as V , keys as K , and the scaling factor as d_k . Therefore, the results of the self-awareness mechanism are calculated using formula 1.

$$Attention(Q, K, V) = softmax(QKT/\sqrt{(d_k)})V, \quad (1)$$

Speech recognition requires the use of neural networks such as transformers, which are designed for speech recognition, usually also called Speech-Transformer. An example of such a neural network and its source code can be found on GitHub [16], this project is called Speech-Transformer, it is developed in Python as a library that can be easily plugged into other projects, so you can simply use it in your own projects.

Transformer-type neural networks for speech recognition transform a sequence of voice features into a corresponding sequence of symbols that together make up words and phrases. The feature sequence, which is longer than the original character sequence, is constructed from

two-dimensional spectrograms with time and frequency dimensions. To be more specific, convolutional neural networks are used to exploit the locality of the spectrograms and mitigate the length mismatch by stepping in time. An example of such a transformer is shown in figure 1.

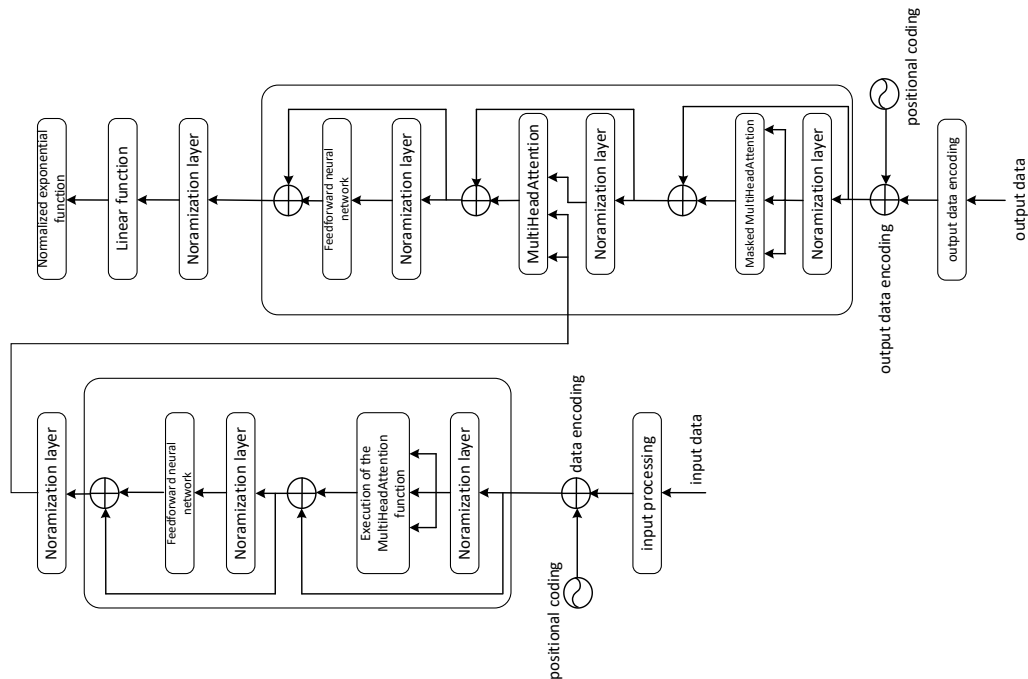


Figure 1: Scheme of the neural network of the transformer type

The queries, keys, and values are extracted from convolutional neural networks and passed to two self-awareness modules. The transformed voice command recognition is evaluated on the WSJ dataset [17] and achieves competitive recognition results with a word error rate of 10.9%, while it requires about 80% less training time than conventional recurrent neural networks, or RNNs for short, and convolutional neural networks, or CNNs.

5. General structure of the system for transforming video materials into the user's language

The method of transforming video materials into the user's language was developed for a hyper-converged platform, which means that it was necessary to adhere to cross-platform, flexible and maximum support for the system on different browsers, as this would facilitate the development and implementation of the system into the platform.

A hyperconverged platform is a large system that consists of many other subsystems, such as a video management system, a live broadcast management system, an event management system, a contract management system, an employee management system, etc. Since the hyperconverged platform and all its modules must run on the Windows operating system at the moment, and in the future must be migrated to the Linux operating system, technologies, libraries and programming languages that support both operating systems were used. To implement the method of transforming video materials into the user's language, which is a module of the hyperconverged platform, Dotnet technology was used for cross-platform compatibility.

The server part is implemented on DotNet technology [18]. This technology was chosen because it provides stability and is suitable for the development of large projects with advanced business logic. In addition, this technology is cross-platform, that is, it can run on many operating systems, namely Windows, Linux and Mac, which greatly facilitates the choice of a server for it. This technology allows developing software of various kinds, for example, for personal computers, mobile devices and websites, APIs games, etc. For the system developed API-type

software to make it easy to transfer data to the client. Also, since other modules of the hyperconverged platform were developed on this technology, its choice will make it easy to connect this module to the platform.

The general architecture of the system for transforming video materials into the user's language shows on Figure 2.

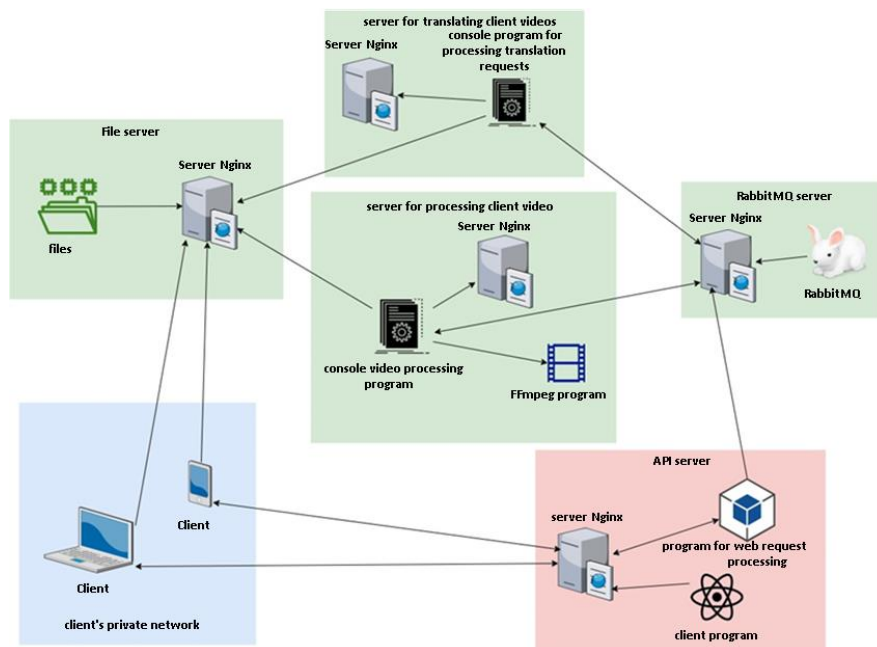


Figure 2: Architecture of the system for transforming video materials into speech user

Software development with the API communication method allows you to connect the developed system anywhere, because this communication method is based on web requests. Thus, the method of transforming video materials into the user's language implemented using the API communication method will work for personal computers, game consoles, phones, for any operating system, as well as for any device that has access to the Internet and supports web requests. Also, the API implementation will allow to connect this system not only to the hyperconverged platform, but also to many other software that can execute web queries, so this system can be used outside the platform, for example, by selling it to other software developers or companies.

The figure shows a server that hosts the server and client parts of the project. This is the basis of the entire solution, since it is through them that the user interacts with the project. It is this project that executes web requests to the server part, which processes incoming requests, reads and writes data to the database, processes data, works with video materials, performs speech recognition, text translation, machine voice over, and produces the result as a JSON object [19].

The client-side implementation of the video transformation method was developed on React [20] because it is very fast and easy to use. This technology also makes it possible to redirect the user to other pages without loading the page, and at the same time, it is possible to show the process of transforming video materials into text in real time. That is, when a user has uploaded his own video to the site and started the process of speech recognition from the video, he will be able to observe this process due to the fact that the text that is recognized will be added to the page sentence by sentence in real time. The same can be applied to other stages, for example, for text translation, you can show in real time how the text is gradually changing and translated into the target language.

Since the first step in the method is the process of separating sound from video material, a separate server is used, since this process takes a lot of computer power, and the presence of a high-performance graphics card on the server significantly speeds up the process, dedicated servers were chosen that do not have a graphics card, but have a good and fast processor.

In order to easily work with video, we chose the "ffmpeg" software [21], but there are no official assemblies for the "dotnet" technology to work with it, so we developed separate Python software [22] that is connected to other projects and used to run "ffmpeg" commands with the correct syntax and in the correct sequence.

Working with video consumes a lot of CPU or video card power, so it is necessary to allocate a separate server for the "ffmpeg" software and install a console program that runs it and indicates which video to open and what to do with it. Otherwise, the server would not withstand the load and would work intermittently.

This python software is console-based, so in order to run it and transmit information to it, you need to use additional software that implements the idea of a message queue. That is, "ffmpeg" is constantly running on the server and constantly listening to the web application port and message queues to instantly respond to a user's request to start processing the video.

Several console programs have been developed to implement a simple and lightweight system for working with video translation. For the systems of text recognition from audio material, text translation, and speech synthesis for translated text, three console-type software applications were developed that wait for a user request and, upon receipt, respond to it, launch the appropriate processes, and provide the result to the user in the form of a JSON object. Each of these programs uses Azure services [23] to quickly and efficiently recognize text, translate it into the target language, and voice the translated text.

To enable users to store video materials, change recognized and translated text, voice acting in different languages, edit and delete them, a file server and a database are used to store the file and all information about it and which video it refers to.

The user also needs to be constantly informed about the state of the system, what is happening with his data, and given the opportunity to observe the transformation of video materials into text and the translation of the processed text into the target language - this is the work of another server to establish a permanent connection with clients and exchange messages in real time.

To establish communication between the server part and other programs on other servers, a server was allocated and the RabbitMQ software was installed on it [24]. Its main task is to send and receive messages from one program to another if they are located, for example, on different servers. It also acts as a message queue.

In general, the work looks like this. The client sends a request to split video and audio into separate files to the server, which has software deployed on it that processes the client's requests, processes them, and sends the user the relevant data in response. This server, in turn, places a message that the video and audio should be split into separate files in the RabbitMQ message queue. In turn, the server that runs the console software for processing video materials must constantly listen to this queue and wait for new requests to be processed. In this way, the system can be scaled horizontally without any changes to the software code, because each server immediately deletes the message from the message queue after receiving it, and thus, if two servers are waiting for a message at the same time, only one will receive it, and the other will wait for new requests further.

Thus, the user can observe text recognition or translation in real time, which makes the system user-friendly.

6. Study of the system functioning

During and after the implementation of the method for transforming video materials into the user's language, we studied the functioning of the implemented system. At first, the system was tested separately from the hyperconverged platform "Iconnect", and then together with the hyperconverged platform to understand whether this implemented system is compatible with the platform.

Since the Dotnet technology was chosen to implement the method of transforming video materials into the user's language, existing libraries that are compatible with this technology can be used to study the functioning of the system.

The xUnit library was used for a modular study of the functioning.

Azure services were used to automate the study of system functioning. After that, every time changes are made to the code, Azure services will receive information about the changes and launch an automated study of the system's functioning.

Figure 3 shows an example of how Azure services work for a project designed to accept user requests, process them, and provide the result to the user in the form of JSON.

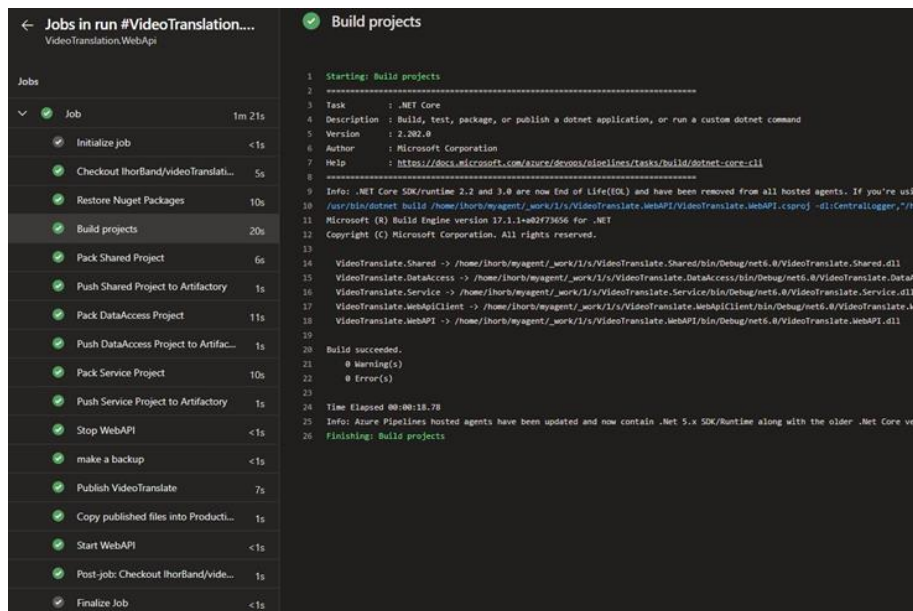


Figure 3: An example of automating the formation of the project for further study of the functioning of the system

Conclusions

This article presents a method for transforming video materials into the listener's language using artificial neural networks. This method plays an important role in the modern world of information technology, where online conferences, video broadcasts, and remote learning are becoming increasingly popular. The transformation process includes several key stages: extracting audio from video, recognizing speech, translating text into the listener's language, and synthesizing voice for an audio file with translated text.

The system ensures high-quality processing at each stage of the process to achieve a satisfactory end result. Recognizing the intonation and emotional tone of speech for accurate translation and reproduction of the text is also an important stage. The use of neural networks significantly improves the efficiency of the speech recognition and translation process, providing more accurate and natural results.

The system was developed on the DotNet platform to ensure cross-platform compatibility and easy integration with other modules of the hyper-converged platform. The system's architecture includes separate servers for processing video, audio, and text data, as well as using Azure cloud services to improve performance and scalability.

This method expands the possibilities for video materials when publishing for a wide audience, regardless of language barriers. It can be applied not only in the entertainment sector but also for educational purposes, promoting the globalization of knowledge and information.

References

- [1] Ning, Yishuang, et al. "A review of deep learning based speech synthesis." Applied Sciences 9.19 (2019): 4050.

- [2] Sarma, Biswajit Dev, and SR Mahadeva Prasanna. "Acoustic-phonetic analysis for speech recognition: A review." *IETE Technical Review* 35.3 (2018): 305-327.
- [3] Trine, Allison, and Brian B. Monson. "Extended high frequencies provide both spectral and temporal information to improve speech-in-speech recognition." *Trends in Hearing* 24 (2020): 2331216520980299.
- [4] Salloum, Said A., Rehan Khan, and Khaled Shaalan. "A survey of semantic analysis approaches." *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*. Springer International Publishing, 2020.
- [5] Livieris, Ioannis E., Emmanuel Pintelas, and Panagiotis Pintelas. "Gender recognition by voice using an improved self-labeled algorithm." *Machine Learning and Knowledge Extraction* 1.1 (2019): 492-503.
- [6] Bhosale Rajkumar, S. "A Holistic Review of Automatic Speech Recognition Systems for Real-time Implementation." *Mathematical Statistician and Engineering Applications* 71.4 (2022): 12341-12359.
- [7] Vieira, Lucas Nunes, Minako O'Hagan, and Carol O'Sullivan. "Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases." *Information, Communication & Society* 24.11 (2021): 1515-1532.
- [8] Viroli, Cinzia, and Geoffrey J. McLachlan. "Deep Gaussian mixture models." *Statistics and Computing* 29 (2019): 43-51.
- [9] Yadav, Munshi, and M. Afshar Alam. "Dynamic time warping (dtw) algorithm in speech: a review." *International Journal of Research in Electronics and Computer Engineering* 6.1 (2018): 524-528.
- [10] Mor, Bhavya, Sunita Garhwal, and Ajay Kumar. "A systematic review of hidden Markov models and their applications." *Archives of computational methods in engineering* 28 (2021): 1429-1448.
- [11] Schmidt, Robin M. "Recurrent neural networks (rnns): A gentle introduction and overview." *arXiv preprint arXiv:1912.05911* (2019).
- [12] Kattenborn, Teja, et al. "Review on Convolutional Neural Networks (CNN) in vegetation remote sensing." *ISPRS journal of photogrammetry and remote sensing* 173 (2021): 24-49.
- [13] Singh, Sushant, and Ausif Mahmood. "The NLP cookbook: modern recipes for transformer based deep learning architectures." *IEEE Access* 9 (2021): 68675-68702.
- [14] iConnectFX™ Community Engagement & Content Sharing Platform. URL: <https://iconnectfx.com/>.
- [15] Peters, Ben, Vlad Niculae, and André FT Martins. "Sparse sequence-to-sequence models." *arXiv preprint arXiv:1905.05702* (2019).
- [16] Speech-Transformer. URL: <https://github.com/sooftware/speech-transformer>
- [17] 32.WSJ language data set. URL: <https://catalog.ldc.upenn.edu/LDC93s6a>
- [18] Build it with .NET. URL: <https://dotnet.microsoft.com/en-us/>
- [19] JSON (JavaScript Object Notation). URL: <https://www.json.org/json-en.html>
- [20] React The library for web and native user interfaces URL: <https://uk.legacy.reactjs.org/>
- [21] A complete, cross-platform solution to record, convert and stream audio and video URL: <https://ffmpeg.org/>
- [22] Python is powerful... and fast. URL: <https://www.python.org/about/>
- [23] Microsoft Azure. URL: <https://azure.microsoft.com>
- [24] RabbitMQ. URL: <https://www.rabbitmq.com/>