COMPUTATIONAL INTELLIGENCE AND MACHINE LEARNING
e-ISSN: 2582-7464

Computational Intelligence and Machine Learning
Vol-4, Issue-2, October 2023

An Overview of Speech-To-Text Conversion

Kartik Aggarwal [1*], Naveen [2]

[1] Dronacharya College of Engineering, Khentawas, Farrukh Nagar, Gurugram, Haryana, India
[2] Assistant Professor, Dronacharya College of Engineering, Khentawas, Farrukh Nagar, Gurugram, Haryana, India
*Corresponding Author Email: kartikaggarwal879@gmail.com

Abstract

As a result of developments in science and technology, an automatic speech-to-text (STT) conversion system has been available. This system converts spoken words into text that can be read visually. People with trouble hearing may use this technology to communicate in other ways, including understanding voice communication and being able to follow directions using their visual abilities. There are instances when seeing something is more powerful than listening to something, particularly in long-distance communication; thus, speech-to-text conversion is crucial in situations like these. One of the fascinating developments to occur in the twenty-first century is the advent of machine learning. It has evolved from its roots in neurology studies conducted in the 1940s into something like artificial intelligence humans have created. Neural networks, a collection of complex structures, are the basis of machine learning. When combined with optimization techniques, these networks mimic the behaviour of neurons in the human brain and allow a computer to learn from its experiences. Here we explore one of many potential uses for such structures - the analysis of vocal performance in an original study. In particular, we dissect voice recognition systems to determine their inner workings.

Keywords

Automated speech-to-text, Machine learning, Natural Language Processing, Neural Networks.

INTRODUCTION

This study examines the present state, development, and significant challenges facing a Speech-to-Text translation accomplished using Artificial Neural Networks. The goal of this research is to improve the accuracy of such translations. This last decade has seen remarkable advancements in artificial intelligence, with practical applications ranging from autonomous vehicles to systems that can analyze neural activations in the brain. The widespread data availability made possible by the rising digitalization of our lives is primarily to blame. Translating human speech into a format that computers can understand and analyze is crucial in today's society, as technological breakthroughs compel us to break down boundaries between people. The capacity to do so is essential in this setting. The study, analysis, and understanding of human speech have the potential to give rise to an unlimited number of practical applications, and this may be one of them. Using neural networks, in particular, building systems that can comprehend and model the human voice with increasing accuracy and reliability is feasible. This has the knock-on effect of making practical tasks such as transcription, translation, and real-time analysis of human language. Computer science and computational linguistics have a subfield called speech recognition that focuses on developing methodologies and technologies that enable computers to recognize spoken language and translate it into text. The discipline of computer science that deals with speech recognition is multidisciplinary.

In today's highly civilized culture, direct human communication is one of the most popular approaches. Applying the various grammatical rules that regulate the creation of phrases, words, and sentences makes it possible for a person to communicate their ideas and thoughts via language [1]. Speech is the primary mode of communication used by humans because it is one of the most natural and effective methods for carrying on conversations and exchanging information and ideas with other people. The distinction between voiced, unvoiced, and silent (VAS/S) sounds in speech is an important acoustic segmentation that may be thought of in this context. It's likely that when this process is applied to a series of sounds known as phonemes, the resulting sound might sound very similar to how the sounds of each letter of the alphabet sound when combined to make human speech [2].

People with extremely high reading and language abilities are the only ones who can access the great majority of the material available on the internet. Linguistic technologies have the potential to provide solutions in the form of standardized user interfaces, which would allow for the widespread transmission of digital content and the facilitation of conversation between speakers of different languages. This technology is necessary in a country like India, which officially recognizes 1652 national languages. The technology known as "speech to text" uses sound picked up by a microphone and transforms it into text that can be seen on a screen [3]. Speech processing is all about rough studying signals and the many different processing processes. Speech processing is a field of research. Various technological tools, such as voice coding, speech synthesis, speech recognition, and speaker identification, are used to carry out this process. Despite the topic's significance, speech recognition was only discussed once [4].

An acoustic signal is sent by a microphone or a phone line, where it is subsequently analyzed by speech recognition algorithms and translated into a preset vocabulary. Electronic circuits or computers are required to decode the linguistic information included in a speech stream. Many other items use this method [5], including security systems, household appliances, mobile phones, ATMs, and laptops. Voice-to-text software performs voice recognition and the translation of spoken languages into text via computational language [6]. There are a few other names for this concept. One of them is computerized speech recognition. Users can read and respond to the content of the audio stream thanks to specific software and technology that can instantaneously transform audio transmissions into text.

MOTIVATION

The purpose of this paper is to compare and contrast the current methods used in deep learning for speech recognition through the study, analysis, and experimentation of different models in light of the enormous societal implications of the application of neural networks and the increasing use of these techniques for the study of the human voice, which seems to improve exponentially as technology advances. To evaluate the advantages and disadvantages of this way of voice modelling, I decided to examine the most current systems and works in the area. In the end, I will also provide the outcomes of several practical experiments in which I trained a model to recognize a specific phrase using convolutional neural networks.

BACKGROUND STUDY

NLP (natural language processing) helps computers understand human speech. Behind the scenes, natural language processing (NLP) analyses the phrases' features and importance and then uses algorithms to extract meanings and provide findings. That is to say, and it makes perfect sense to automate several tasks using human language [9].

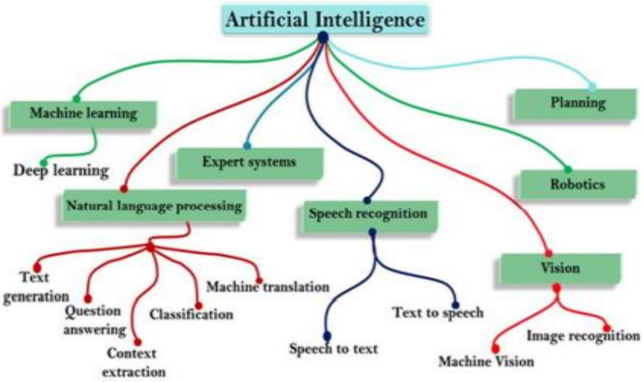

Figure 1. Artificial intelligence's use of natural language processing [11]

The field of study known as Natural Language Processing (NLP) looks at how computers can translate and comprehend human speech [10]. Using natural language processing (NLP), computers can translate text, extract keywords, classify topics, and more from written or spoken input [11].

1. The greater the amount of pertinent information in the questions, the higher the probability that the answer will be correct.
2. In a matter of seconds, users may obtain a direct response to their questions regarding any issue.
3. One of its main selling points is how easy it is to implement.
4. The usage of the software is more cost-effective than employing a person. It's possible that a machine could do the same tasks in half the time it would take a person.
5. Naturally language processing (NLP) answers questions asked in that format.
6. Benefit from more language-based data collection without fatigue and do it more objectively and consistently than a human.
7. The NLP technique facilitates computer-human interaction by allowing it to communicate in its native tongue and by providing a means of quantifying other linguistic processes.

Recognition of Speech with the Use of Recurrent Neural Networks

The subject of this study, which focuses on speech recognition using recurrent neural networks rather than a CNN, piqued my interest. We now have a better knowledge of the differences between the two types of networks and how each functions as a result of this research. The main difference between the two models is that convolutional neural networks cannot process sequential input. This represents the distinction between the two models. This is because the network evaluates each piece of information by considering just the present state of each neuron, unaffected by previous inputs. This feature is also available in recurrent neural networks. Each neuron has a type of "memory" that permits it to review and process each input while considering some of the information from the most recent input it had previously processed. This technique is required for processing sequential data since successive items of information nearly always have a relationship between them that must be considered. For more sophisticated voice recognition applications, such as those that aim to handle more words or complex phrases, recurrent neural networks are required [43]. A convolutional neural network performs well for tasks such as audio recognition for single and few words.

Neural Networks for the translation of spoken language into text

To translate spoken English and Russian into text, studies including deep recurrent neural networks equipped with LSTM (long short-term memory) are now being carried out as part of this research. The data source for the training and validating of a deep neural network was English audiobooks. The training was carried out on the network. In order to compile the neural network learning dataset, first an English audiobook and then a Russian text was segmented into sentences, and then features were extracted from audio files. The two texts were used as the basis for the neural network learning dataset. TensorFlow is a machine learning library containing a set of algorithms that can be used to train and assess a neural network. These algorithms may be accessed via the TensorFlow library. The calculations were carried out on a server that was running Linux and had two NVIDIA video cards in it. The server had a high performance so that it could handle the workload. Two distinct models were taught: models for translating speech into text and models for translating text into speech. Both kinds of models were learned. As a result of the study, it became abundantly obvious that models built via deep neural networks can deliver reliable results when applied to the activity of automated language translation [44].

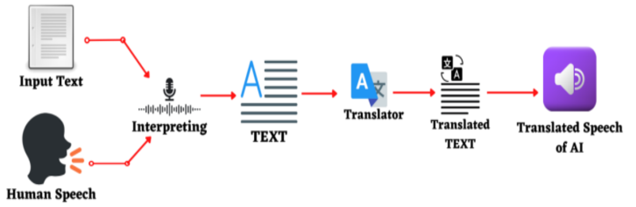

Figure 2. Translation of Voice and Text [45]

The Challenges Faced by Natural Language Processing (NLP)

Natural Language Processing (NLP) is a robust technique that offers enormous advantages; nonetheless, there are still several limits and issues associated with Natural Language Processing:

- Words and phrases that are contextually relevant as well as homonyms
- Synonyms
- Irony, as well as sarcasm
- Ambiguity
- Errors in text or speaking
- Slang and colloquialisms in everyday use
- Terminology unique to a particular domain
- Languages with limited access to resources
- A lack of investment in research and development

LITERATURE REVIEW

According to the study, new technologies are needed to preserve one of the world's oldest languages, Tamil. The study focuses on indigenous design difficulties for these systems [7]. Our technique decreases Librispeech ASR WER by 14% to 19%. We also tested its influence on spoken language understanding (SLU) and saw a 2.5% to 2.8% F1 improvement on the SNIPS slot-filling test [8].

Based on the findings, it seems plausible to create systems that convert voice to meaning and meaning to speech without text [9]. Release the 11.5-hour transcripts. The best model, with a WER of 3.10% [10], uses the Kaldi architecture, which includes a Deep Neural Network. We utilize the Multilingual Speech Translation Corpus to test a meta-learning approach for ST tasks (MuST-C). Our solution surpasses existing transfer learning methods by 9.18 BLEU points for En-De ST issues [11]. This program's cloud Firebase handles user authentication. Registered users may utilize voice-to-text. Hidden Markov models that use audio inputs are used to assess application performance. The programme has 86.8% accuracy, 86.8% f1, 91.5% precision, and 95.4% recall [12]. ML and DL will be utilized to build a Kannada and Hindi POS tagger. Experiments used 3 million distinct Kannada and Hindi words. 13 BIS POS tags. This technique uses speech embeddings from ASR. Speech-only tests are 15% more accurate than ASR-based text tests when using genuine transcripts [14]. FastText characteristics showed a 95.4% accuracy for recognizing insult remarks and a 93.9% accuracy for identifying hate comments. Our research found Chadian hate speech online [15]. The study examines spelling mistakes and implementation. This study develops a gender- and speaker-independent continuous speech-to-text system [16]. The Support Vector Machine outperformed 92% of previous Afan Oromo hate speech classifiers. These Afan Oromo hate speech writings are online [17]. A Bengali spell checker improves system performance [18]. Filipino Speech to Text [19] uses convolutional neural networks to turn audio recordings into three-dimensional word representations. Deep learning models offer more promise than conventional ones. We'll test this further [20]. After then, the sounds are transformed into text using the speech-to-text module of IBM Watson. Once text chunks have been retrieved, they are combined before being sent to the Google Machine Translation API to convert into Indian. Following the completion of the translation, a TTS system is necessary to turn the text into audio. There are no open-source TTS solutions available for India's regional languages. Now in production is the Flite engine, an improved and more portable variant of Professor Alan Black's Festival engine (CMU). TTS is responsible for producing sounds from translated text within this flite engine. The accuracy of the programme may go as high as 91% for a single video and is often around 79%. The audio's naturalness is evaluated by contrasting it with language often used in everyday situations. This is a helpful tool for those who cannot read and those who do not read to improve their understanding of their native language. The long-term goal of this software is to facilitate global communication,which would make it possible for people of any language to have conversations [21]. This study provides an overview of transcription and speech synthesis systems that are based on deep learning. It can convert voice into text, text into speech, and speech into text that can be spoken. Experiments involving voice-to-text conversion and speech synthesis may be found in our collection. This study also analyses the experimental outcomes of applying two state-of-the-art pre-trained models to various test conditions and comparing the results. AI may power future video conferencing systems [22]. This study provides an overview of transcription and speech synthesis systems that are based on deep learning. It can convert voice into text, text into speech, and speech into text that can be spoken. Experiments involving voice-to-text conversion and speech synthesis may be found in our collection. In addition, the experimental findings of two cutting-edge pre-trained models are also scrutinized. AI may power future video conferencing systems [23].

The suggested system has two stages: the first is data preprocessing, and the second is gesture modelling. The first thing that we do is combine the text and the audio into a single input vector. After that, we use representation learning to preliminary process the 3D motion data. Only 15 of the joints in the upper body were chosen to be represented by a vector after being educated by the DAEs model. In the second step of the process, the encoder-decoder bidirectional LSTM architecture is used to construct the gesture model. Acoustic speech and text are converted into DAE output by this approach. 3D motion output video production occurs due to decoding the DAE representation. The system was tested using a held-out dataset and compared to nine other techniques, including five individuals, two baseline systems, and two natural movements (real video). The human likeness and appropriateness of the creations were evaluated with the help of the crowd. Both strategies prioritize humans. The results of both of these subjective evaluations place our proposed approach in the middle [24]. CNN and RNN are examples of deep learning algorithms. When combined, these two approaches make it possible to recognize sign language automatically. Both of these tactics improve the system's performance, which now stands at an estimated 92.4% accuracy on dynamic hand movements when evaluated on most available datasets. This kind of device might enable a person to give a presentation or participate in a video conference on a platform used in business or education while simultaneously displaying an image- or video-based representation of sign language in real time. After recognizing sign language, the system will convert it to text, then flip the text to speech using a Python Text-to-Speech API. This well-developed architecture has the potential to solve several communication problems [25].

A semi-supervised LDA algorithm should be used when combining the information from the slide text and the audio transcript. By using seed words shown on video slides, one of our other goals is to enhance the learning capabilities of the model. The system is taught to read video transcriptions based on seed words. We put the multimodal indexing method through its paces on five hundred films sourced from Coursera, NPTEL, and KLETU (KLE Technological University). Multimodal fusion enhances the F-score by 44.49% compared to unimodal indexing [26]. The following article will offer a system that instantly transforms participants' spoken words in a videoconference into text. The technique that has been provided is reliant on JavaScript-based browser APIs. These APIs are referred to as WebRTC and Web Speech APIs. These application programming interfaces allow developers to create programmes that facilitate video conferencing, messaging, sharing desktops, and transferring data from one browser to another. Developers can do real-time speech-to-text conversions using the Web Voice API. The adoption of OpenVidu, an open-source videoconferencing system that is based on WebRTC, is the solution that is advocated. After that, you should link the front end of OpenVidu with the Web Speech API [27].

A combination of audio, video, and emotional summaries is used in natural language processing. Tokenization, phrase segmentation, lemmatization, stemming, and abstractive summarisation are some of the methods that fall under this category. A meaningful and accurate film description may be obtained via the use of a video summary. When accurate descriptions are accurate, locating material that matches them is much simpler. [28] The participants in the research had an 87% confidence that the text properly reflected the video. Visually impaired individuals may improve their lives by having access to video-to-Braille transcription at any time and in any location. For those who are hard of hearing or visually handicapped, we have developed software that can transcribe video into Braille. It would be of substantial benefit to implement a system for visually impaired people that enables real-time voice-to-Braille transcription, text-to-speech conversion, text-to-Braille conversion, and video-to-Braille conversion, as well as flexible command work [29]. The proposed method uses an ad hoc phase of Named Entity Recognition and Disambiguation (NERD) to analyze the video text and audio transcripts provided by the teacher (for example, the content of the slide and the note written on the whiteboard). NERD makes use of a methodology that is unique to the realm of supervised classification. Not only does it take into account text similarity metrics, but it also takes into account the entity's semantic pertinence to the primary topic that is covered in the video lectures in order to choose the most prominent entities in the knowledge base that correlate to the video content. GERBIL's tools were used to verify the suggested system's performance. The preliminary findings indicate that the presented method is successful [30]. This compilation of movies includes extensive audio and text annotations that are tied to the sequence of events. The content descriptions are more pertinent than the conversation and are more extensive than the previous explanations, which either needed to be more superficial or informative. It is possible to train retrieval and event localization algorithms through their paces and test them on the QuerYD dataset. This example shows the usefulness of QuerYD. We have high hopes that QuerYD will inspire more research on understanding videos via the use of written and spoken natural language [31]. This makes our solution generative and allows it to be utilized for various "video+x to text" tasks without creating new network heads. Our single-architecture method beats the state-of-the-art in captioning, question-answering, and audio-visual scene-aware conversation [32]. The structure is essential and effective. Suicide may be prevented by identifying suicidal thoughts early. Because of internet communication and human connection, people may instantaneously share their challenges and sentiments, which can help identify suicidal ideation. These parameters vary from person to person [33]. Speech-to-text transcribes criminal news videos.

The Thai words Segmentation module then tokenizes the passage into words. Apply the Gestures sequence construction module to the gesture videos. This study also analyzed the National Association of the Deaf in Thailand's results for its benefits and future adjustments [34]. This study provides a consistent, optimized paradigm for speech summarization. We adopt limited self-attention from text-based models to speech models to alleviate memory and processing restrictions. We demonstrate that the proposed model can directly summarise speech using How-2 as training data. The end-to-end model is 3 ROUGE points better than the cascaded model. We also include interpreting spoken language, which involves predicting ideas based on voice inputs, and demonstrate that the end-to-end model beats the cascade model by four absolute F-1 [35]. Considering spoken language comprehension does this.

Speech-to-text is used to transcribe and analyze an educational film. To organize microlearning videos into learning categories, a three-tiered architecture is provided. "MVR-CLS" can classify microlearning videos more granularly than previous studies. The classification result is compared to OER information to evaluate the proposed technique. The classification result may help match content suggestions to learner preferences, improving future content recommendations [36]. In addition to the SpeechRecognition library, the Google SpeechToText API is utilized to discern speakers from silences in transcriptions. BERT2BERT, a BETO model pre-trained on Hugging Face Transformers news summaries [37], summarises for us. In this study, we investigate the Semantic Association Network (SAN) for Video Corpus Instant Retrieval (VCMR), which localizes the temporal point in a collection of movies that most closely fits the text query. VCMR is an acronym that stands for Video Corpus Instant Retrieval. Videos that include text queries and subtitles are required to take advantage of standard semantics that is created from several input modes. For the purpose of this cooperation, SAN associates standard semantics both inside the same modality (through the process of intra-semantic association) and across modalities (through the process of inter-semantic association) (MSA). On the TVR and DiDeMo benchmark datasets, SAN performs better than the current state of the art. The efficacy of the method has been shown by a number of extensive ablation experiments as well as qualitative reviews [38]. With only one voice command, the user may watch video lessons and have essential guidelines read aloud to them. The system analyses the acoustic fingerprint of the spoken command and compares it to the relevant moment in the movie. The index table, which connects transcription to the occurrence, is kept up to date by the system that has been explained. The interactive play control looks for the keyword and then begins playing the multimedia file from the specified location. [39] Completing and simplifying the system is achieved by the integration of efficient transcription, storage, and retrieval.

A PFE module that extracts POS trends predicts POS tags, and fuses visual attributes is what we propose. This module will employ a variety of filters. In addition to this, we suggest a visual-dynamically aware (VDA) module that may dynamically change word mapping and provide visual information to local features. The fusion characteristics provide visual information that may be used to generate accurate words, and the anticipated POS tags offer assistance with decoding, which helps to produce a syntax structure that is more standardized and unified. This is achieved by generating correct words and providing direction for decoding. As shown by several MSVD, MSR-VTT, and VATEX tests, our approach is superior to the most recent and cutting-edge methods for solving BLEU-4, ROUGE-L, METEOR, and CIDEr [40].

The first step in the consultation process is to voice-record the consultation data, followed by recording the prescription using Speech to Text and Microsoft Azure's cognitive speech service. It will build a medical system based on a hybrid reality and use technology from the metaverse in order to provide more online interactive services. In further research, the goal is to find a way to combine a virtual reality rehabilitation training system with telemedicine so that it can provide a more all-encompassing service [41].

The proposed framework makes an effort to use audio information even if audio inputs are not supplied while making inferences. By using linked visual features, the proposed VAM can inscribe the audio data of short clips into a memory network. In order to do this, the VAM will mix the lip-video key with the audio value. The audio value memory is responsible for making an impression of the audio feature, while the lip-video key memory is responsible for remembering its location. When this is accomplished, the VAM will be able to access the memory in order to make use of the much auditory information. Experiments demonstrate that the strategy that is advocated produces state-of-the-art VSR on both the word level and the sentence level. In addition to this, we make sure that the learned representations of the VAM include all of the relevant information from the VSR [42].

Table 1: The many different models available for converting speech to text

S.No.	Method	Advantage	Disadvantage
1.	Linear Predictive Coding (LPC)	**1.** LPC is a Static method that is utilized for the extraction of features. **2.** The basic idea behind linear predictive coding (LPC) is that it can use a speech sample to create a linear combination by mixing previous voice samples.	Relies on spectral analysis with a predetermined resolution and a purely evaluative frequency scale.

S.No.	Method	Advantage	Disadvantage
		3. The audio signal is split into a series of frames called N, and then the information inside these frames is transformed into text.	
2.	Mel-Frequency Cestrum Coefficient (MFCC)	• Multi-Frequency Concatenation (MFCC) is yet another method that involves extracting signal characteristics using filter banks. • In order to convert STT data, the method utilizes processes such as framing, windowing, and the discrete Fourier transform.	The challenge with MFCC is that it has to be normalized since the values it generates could be more effective in the presence of additive sounds or their surroundings.
3.	Dynamic Time Wrapping	• Using dynamic programming, the DTW method is used to determine the analogies between two-time series occurrences that differ in the pace at which they occur. It does this by iterating through the pairwise sequence of feature vectors and looking for a plausible match between them.	The issue emerges when choosing the reference template to use in contrasting the events that have occurred throughout time.
4.	Hidden Markov Model	• HMM is a statistical model used in the STT conversion process. • Because HMMs have their own structure and are capable of self-learning, they are very helpful for STT conversion.	When using this technique, the speech signal is interpreted as a static signal or a static signal for a brief period. HMM is a serial process.
5.	Neural Network	**1.** A neural network is both a statistical model and a representation of that model in the form of a graph. **2.** In order to complete state transactions, neural networks make use of connection functions' values as well as connection strengths.	Here, in the ANN-parallel neural network model.
6.	Hybrid Approach	**1.** The suggested hybrid method is utilized for converting voice to text because speech frequencies are parallel, but syllable series and word series are serial. This is due to the fact that words and syllables are broken down into series. **2.** This demonstrates that any of these approaches may be helpful in a variety of settings. Techniques from Hidden Markov Models (HMM) and Neural Networks are combined in this implementation. **3.** Since neural networks have shown strong performance in analyzing the probability from parallel voice input, and since Markov models can make use of the phoneme observation probabilities that neural networks supply to generate the probable phoneme sequence or word.	

Obstacles encountered by systems of automated speech recognition (ASR)

- A lack of knowledge of other languages.
- Sounds in the background and in the periphery.
- The unreliable nature of the data provided by ASR.
- Costs and strategies for deployment.

CONCLUSION

Several aspects of unsupervised speech recognition were examined throughout our research, including unsupervised sub-word modelling, spoken word embeddings, unsupervised term discovery, full-coverage segmentation, and cross-modal alignment. It is common knowledge that translation is a handy tool for communication; thus, if you run a business that operates on a worldwide scale and does not use the benefits that translation services provide, you run the risk of losing out on prospective customers. Including translation services in fundamental aspects of one's business, such as providing one's website in a variety of languages or conducting marketing activities in a number of languages, is a simple and effective way to wow one's foreign clients and make them feel as if they are Because of this, your company's reach to a wider variety of individuals all over the globe will even be increased. We are aware that employing artificial intelligence alone makes machine translation the fastest method of translating text from one language to another. On the other hand, a human translation requires genuine critical thinking in the form of one or more translators manually doing a translation. The quality of translations generated by machine translation is advancing at a pace that is quicker than its previous rate of improvement. Even if the level of fluency is increasing, it is still vital to have human translators examine the finished product since there is still a possibility that errors may be made in the translation.

REFERENCES

[1] Ö. B. Mercan, U. Özdil and Ş. Ozan, "Increasing Performance in Turkish by Finetuning of Multilingual Speech-to-Text Model," 2022 30th Signal Processing and Communications Applications Conference (SIU), 2022, pp. 1-4, doi: 10.1109/SIU55565.2022.9864728.

[2] T. Kano, S. Sakti and S. Nakamura, "End-to-End Speech Translation With Transcoding by Multi-Task Learning for Distant Language Pairs," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1342-1355, 2020, doi: 10.1109/TASLP.2020.2986886.

[3] J. Kim, M. Kumar, D. Gowda, A. Garg and C. Kim, "Semi-Supervised Transfer Learning for Language Expansion of End-to-End Speech Recognition Models to Low-Resource Languages," 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2021, pp. 984-988, doi: 10.1109/ASRU51503.2021.9688019.

[4] T. Wang, J. Yi, R. Fu, J. Tao and Z. Wen, "CampNet: Context-Aware Mask Prediction for End-to-End Text-Based Speech Editing," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 2241-2254, 2022, doi: 10.1109/TASLP.2022.3190717.

[5] J. Iranzo-Sánchez et al., "Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 8229-8233, doi: 10.1109/ICASSP40776.2020.9054626.

[6] P. Kaur, S. Ramu, S. Panchakshari and N. Krupa, "Conversion of Hindi Braille to Speech using Image and Speech Processing," 2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), 2020, pp. 1-6, doi: 10.1109/UPCON50219.2020.9376566.

[7] D. Pubadi et al., "A focus on codemixing and codeswitching in Tamil speech to text," 2020 8th International Conference in Software Engineering Research and Innovation (CONISOFT), 2020, pp. 154-165, doi: 10.1109/CONISOFT 50191.2020.00031.

[8] W. Wang et al., "Optimizing Alignment of Speech and Language Latent Spaces for End-To-End Speech Recognition and Understanding," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 7802-7806, doi: 10.1109/ICASSP43922. 2022.9747760.

[9] O. Scharenborg et al., "Speech Technology for Unwritten Languages," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 964-975, 2020, doi: 10.1109/TASLP.2020.2973896.

[10] D. Ungureanu, M. Badeanu, G. -C. Marica, M. Dascalu and D. I. Tufis, "Establishing a Baseline of Romanian Speech-to-Text Models," 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2021, pp. 132-138, doi: 10.1109/SpeD53181.2021.9587345.

[11] S. Indurthi et al., "End-end Speech-to-Text Translation with Modality Agnostic Meta-Learning," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7904-7908, doi: 10.1109/ICASSP40776.2020.9054759.

[12] R. Kumar, M. Gupta and S. R. Sapra, "Speech to text Community Application using Natural Language Processing," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), 2021, pp. 1-6, doi: 10.1109/ISCON52037.2021.9702428.

[13] V. Advaith, A. Shivkumar and B. S. Sowmya Lakshmi, "Parts of Speech Tagging for Kannada and Hindi Languages using ML and DL models," 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2022, pp. 1-5, doi: 10.1109/CONECCT55679. 2022.9865745.

[14] L. Sarı, S. Thomas and M. Hasegawa-Johnson, "Training Spoken Language Understanding Systems with Non-Parallel Speech and Text," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 8109-8113, doi: 10.1109/ICASSP40776. 2020.9054664.

[15] M. S. Adoum Sanoussi, C. Xiaohua, G. K. Agordzo, M. L. Guindo, A. M. Al Omari and B. M. Issa, "Detection of Hate Speech Texts Using Machine Learning Algorithm," 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), 2022, pp. 0266-0273, doi: 10.1109/CCWC54503.2022.9720792.

[16] S. Saha and Asaduzzaman, "Development of a Bangla Speech to Text Conversion System Using Deep Learning," 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 2021, pp. 1-7, doi: 10.1109/ICIEVicIVPR52578.2021.9564209.

[17] N. B. Defersha, K. Kekeba and K. Kaliyaperumal, "Tuning Hyperparameters of Machine Learning Methods for Afan Oromo Hate Speech Text Detection for Social Media," 2021 4th International Conference on Computing and Communications Technologies (ICCCT), 2021, pp. 596-604, doi: 10.1109/ICCCT53315.2021.9711850.

[18] H. M. M. Hasan, M. A. Islam, M. T. Hasan, M. A. Hasan, S. I. Rumman and M. N. Shakib, "A Spell-checker Integrated Machine Learning Based Solution for Speech to Text Conversion," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 1124-1130, doi: 10.1109/ICSSIT48917.2020.9214205.

[19] S. G. E. Brucal et al., "Filipino Speech to Text System using Convolutional Neural Network," 2021 Fifth World Conference on Smart Trends in Systems Security and Sustainability (WorldS4), 2021, pp. 176-181, doi: 10.1109/WorldS451998.2021.9513991.

[20] D. Escobar-Grisales et al., "Colombian Dialect Recognition Based on Information Extracted from Speech and Text Signals," 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2021, pp. 556-563, doi: 10.1109/ASRU51503.2021.9687890.

[21] S. K. Pulipaka, C. K. Kasaraneni, V. N. Sandeep Vemulapalli

and S. S. Mourya Kosaraju, "Machine Translation of English Videos to Indian Regional Languages using Open Innovation," 2019 IEEE International Symposium on Technology and Society (ISTAS), 2019, pp. 1-7, doi: 10.1109/ISTAS48451.2019.8937988.

[22] S. Tanberk, V. Dağlı and M. K. Gürkan, "Deep Learning for Videoconferencing: A Brief Examination of Speech to Text and Speech Synthesis," 2021 6th International Conference on Computer Science and Engineering (UBMK), 2021, pp. 506-511, doi: 10.1109/UBMK52708.2021.9558954.

[23] S. Tanberk, V. Dağlı and M. K. Gürkan, "Deep Learning for Videoconferencing: A Brief Examination of Speech to Text and Speech Synthesis," 2021 6th International Conference on Computer Science and Engineering (UBMK), 2021, pp. 506-511, doi: 10.1109/UBMK52708.2021.9558954.

[24] Thangthai, K. Thangthai, A. Namsanit, S. Thatphithakkul and S. Saychum, "Speech Gesture Generation from Acoustic and Textual Information using LSTMs," 2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2021, pp. 718-723, doi: 10.1109/ECTI-CON51831.2021.9454931.

[25] Sonare, A. Padgal, Y. Gaikwad and A. Patil, "Video-Based Sign Language Translation System Using Machine Learning," 2021 2nd International Conference for Emerging Technology (INCET), 2021, pp. 1-4, doi: 10.1109/INCET51464.2021.9456176.

[26] M. Husain and S. M. Meena, "Multimodal Fusion of Speech and Text using Semi-supervised LDA for Indexing Lecture Videos," 2019 National Conference on Communications (NCC), 2019, pp. 1-6, doi: 10.1109/NCC.2019.8732253.

[27] S. Eltenahy, N. Fayez, M. Obayya and F. Khalifa, "Conversion of Videoconference Speech into Text based on WebRTC and Web Speech APIs," 2021 International Telecommunications Conference (ITC-Egypt), 2021, pp. 1-4, doi: 10.1109/ITC-Egypt52936.2021.9513968.

[28] Emad, F. Bassel, M. Refaat, M. Abdelhamed, N. Shorim and A. AbdelRaouf, "Automatic Video summarisation with Timestamps using natural language processing text fusion," 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 2021, pp. 0060-0066, doi: 10.1109/CCWC51732.2021.9376115.

[29] Verma, A. Rai, H. Yadav, V. Rastogi and S. Satija, "Video To Braille Transcription For Visually Impaired People," 2021 International Conference on Simulation, Automation & Smart Manufacturing (SASM), 2021, pp. 1-3, doi: 10.1109/SASM51857.2021.9841184.

[30] L. Cagliero, L. Canale and L. Farinetti, "VISA: A Supervised Approach to Indexing Video Lectures with Semantic Annotations," 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), 2019, pp. 226-235, doi: 10.1109/COMPSAC.2019.00041.

[31] -M. Oncescu, J. F. Henriques, Y. Liu, A. Zisserman and S. Albanie, "QUERYD: A Video Dataset with High-Quality Text and Audio Narrations," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 2265-2269, doi: 10.1109/ICASSP39728.2021.9414640.

[32] X. Lin, G. Bertasius, J. Wang, S. -F. Chang, D. Parikh and L. Torresani, "VX2TEXT: End-to-End Learning of Video-Based Text Generation From Multimodal Inputs," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7001-7011, doi: 10.1109/CVPR46437.2021.00693.

[33] G. Chordia, Y. Mehta, S. Nayan, H. Soni, A. Gupta and J. K, "Suicidal Prediction Using Video, Audio, And Text Analysis," 2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS), 2022, pp. 40-45, doi: 10.1109/ICPS55917.2022.00015.

[34] N. Namyang, J. Lumpaolertwilai and S. Phimoltares, "Speech-to-Thai Sign Language Conversion for Thai Deaf: A Case Study of Crime News," 2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2019, pp. 115-120, doi: 10.1109/JCSSE.2019.8864203.

[35] R. Sharma, S. Palaskar, A. W. Black and F. Metze, "End-to-End Speech Summarisation Using Restricted Self-Attention," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8072-8076, doi: 10.1109/ICASSP43922.2022.9747320.

[36] S. -Y. Chong, F. -F. Chua and T. -Y. Lim, "MVR-CLS: An Automated Approach for Effective Classification of Microlearning Video Resources," 2022 International Conference on Advanced Learning Technologies (ICALT), 2022, pp. 74-76, doi: 10.1109/ICALT55010.2022.00029.

[37] F. J. Lozano, D. Alcázar and A. Díaz, "Web-based application for generation of video-interview summaries using neural networks," 2022 17th Iberian Conference on Information Systems and Technologies (CISTI), 2022, pp. 1-3, doi: 10.23919/CISTI54924.2022.9820457.

[38] D. Kim, S. Yoon, J. W. Hong and C. D. Yoo, "Semantic Association Network for Video Corpus Moment Retrieval," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 1720-1724, doi: 10.1109/ICASSP43922.2022.9747523.

[39] G. N, V. P and P. S, "Interactive Audio Indexing and Speech Recognition based Navigation Assist Tool for Tutoring Videos," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022, pp. 1678-1682, doi: 10.1109/ICSCDS53736.2022.9760784.

[40] L. Wang, H. Li, H. Qiu, Q. Wu, F. Meng and K. N. Ngan, "POS-Trends Dynamic-Aware Model for Video Caption," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 7, pp. 4751-4764, July 2022, doi: 10.1109/TCSVT.2021.3131721.

[41] P. -J. Lin, B. -C. Tsai and Y. -W. Tsai, "Telemedicine System Based on Mixed Reality and Cognitive Speech Service Technologies," 2022 IEEE 4th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), 2022, pp. 241-244, doi: 10.1109/ECBIOS54627.2022.9944986.

[42] M. Kim, J. Hong, S. J. Park and Y. M. Ro, "CroMM-VSR: Cross-Modal Memory Augmented Visual Speech Recognition," in IEEE Transactions on Multimedia, vol. 24, pp. 4342-4355, 2022, doi: 10.1109/TMM.2021.3115626.

[43] Amberkar, P. Awasarmol, G. Deshmukh and P. Dave, "Speech Recognition using Recurrent Neural Networks," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 2018, pp. 1-4, doi: 10.1109/ICCTCT.2018.8551185.

[44] R. F. Gibadullin, M. Y. Perukhin and A. V. Ilin, "Speech Recognition and Machine Translation Using Neural Networks," 2021 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM),

2021, pp. 398-403, doi: 10.1109/ICIEAM51226.2021. 9446474.

[45] Adiba Noor , Krish Chopra , Ayush Minj , Priyesh Kumar, Kauleshwar Prasad, Dinesh Kumar Bhawnani ," Voice & Text Translator", Research and Applications of Web Development and Design Volume 5 Issue 1 , 2022 , pp. -5