Development of Multilingual Speech Recognition and Translation Technologies for Communication and Interaction

Ali A. AL-Bakhrani[1,4](✉), Gehad Abdullah Amran[2], Aymen M. Al-Hejri[4,5], S. R. Chavan[3], Ramesh Manza[3], and Sunil Nimbhore[3]

[1] Department of Computer Science, Technique Leaders College, Sana'a, Yemen
albakhrani2017@gmail.com

[2] Department of Management Science and Engineering, Dalian University of Technology, Dalian, Liaoning 116024, China

[3] Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India

[4] Faculty of Administrative and Computer Sciences, Albaydha University, Albaydha, Yemen

[5] School of Computational Sciences, Swami Ramanand Teerth Marathwada University, Nanded, Maharashtra, India

Abstract. In this study, we find a solution to the problem of recognizing the source language and translating it into the selected target language. This interface is designed to convert the voice or speech into any selected source text, convert it into the targeted text, and save it into wave files. This interface, which in turn solves many problems, including in the field of education and society, can be used in day-to-day life. We have worked on building a software project that solves the problem, as it relies on deep learning techniques in speech recognition. Building the application depends on several main parts: speech recognition, verification of the speaker's language, conversion of speech to text, translation of speech into any language, and conversion into any language. The text of the speaker or translator into voice also allows saving speech in a pdf file and supports translating entire files, as this application has been programmed using the Python programming language.

Keywords: TTS · STT · DNN · Speech Recognition · Translation · Python

1 Introduction

People use speech to communicate with each other, which is the most natural and efficient way to exchange information [1]. From this conclusion, it follows that the next technological advancement will be natural language voice recognition for human-computer interaction. Speech recognition is a branch of computer science and computational linguistics that deals with methodology and technology development for computer speech recognition. Automatic speech recognition (ASR), computer speech recognition, or speech-to-text is sometimes known as ASR, computer speech recognition, or speech

R. Manza et al. (Eds.): ACVAIT 2022, AISR 176, pp. 711–723, 2023.
https://doi.org/10.2991/978-94-6463-196-8_54

recognition (STT) [1]. It integrates technology into the domains of linguistics, computer science, and computer engineering. For voice recognition systems to function correctly, they must be “trained” (sometimes called “enrolled”). By analyzing a person’s voice, the system’s recognition of that person’s speech is more accurate because of the fine-tuning method used. Conversational spoken sentences are quickly translated and spoken aloud in a second language through speech translation [1]. System phrase translation varies from phrase translation because it does not translate phrases that are fixed and finite, but rather anything and everything that can be used in a sentence. Speech translation technology allows people to converse with each other, regardless of their native language [8]. Thus, it is of considerable benefit to humanity in terms of science, intercultural dialogue, and business worldwide. People naturally assume that speaking will be the way communication with computers is conducted. A computer capable of speaking and understanding its native language.

Machine reformation of speech uses a sequence of words and attempts to find the best possible match for the provided speech signal [13]. Virtual reality, multimedia searches, and flight check-in agents are some of the applications for knowledge creation, including informative and on-site accommodations, interpreters, and natural language comprehension. To expand the concept further and make the best use of these technologies, we combined them into one application, which was programmed in the Python language, which supports the user interface with the addition of new features, which serve precisely in the field of education and scientific conferences, and we relied on artificial intelligence techniques [14], which Google translators relied on. It works to solve a problem in education, which in turn provides educational material to the fullest and with high efficiency. For example, the lecturer speaks English, and the students have different languages, such as some of the students speaking Arabic and some of the students speaking Hindi. The student can choose his language, and the application also saves the entire lecture in a pdf file and downloads a file in the speaker’s language and another file in the student’s language, which helps the student write all observations and focus on the teacher.

2 Related Work

Few surveys have been conducted on voice recognition in this region. Consider, for instance, an evaluation of voice recognition and feedforward networks aided by discriminatively trained networks. The primary purpose of the review was to identify publications that used several processing steps prior to HMM-based word sequence decoding. A number of computational approaches, some of which provide significant improvements in very short vocabulary problems while simultaneously increasing the signal-to-noise ratio (SNR) in very large vocabulary tasks, were discussed in [1]. In addition, the construction techniques outlined offered step-by-step instructions for building structures incorporating many layers, which were built from several layers of MLPs with a high number of hidden layers. This research ultimately concluded that although deep processing structures can develop in this genre, several aspects, such as layer width, have a significant impact on it. Deep neural networks that employ several hidden layers and are trained with new approaches are now being utilized. This is an excellent summation of the results of four separate research groups that cooperate to conclude that feedforward neural networks are best equipped to handle both HMM states and HMM coefficients. In lieu of utilizing standard HMMs and GMMs for acoustic modeling in speech recognition, this alternative was investigated. In a study that has been in progress for several years, deep neural networks that contain many hidden layers and are trained by innovative approaches have demonstrated significant performance gains over GMMs (HMMs) on voice recognition benchmarks [2]. This study states that deep neural networks for acoustic models of speech recognition have an extended history. The overview report concentrates on how deep learning techniques can be further improved. These improved approaches include improved network design and activation functions, improved optimization methods, and new methods for finding neural network parameters. From the overview, we can see that acoustic models that employ deep neural networks (which may or may not use GMMs) are making great progress. Other signal-processing applications, such as voice recognition, may also benefit from using similar acoustic models [3]. To compare Microsoft's recent progress in voice recognition in 2009, a summary was created using deep learning. This study sought to understand recent advancements in voice recognition by investigating the capabilities and limits of deep learning in the field. Microsoft supplied samples from their latest research to facilitate the incorporation of deep learning algorithms for speech-related applications. The incorporation of applications in the field of speech-related technology includes feature extraction, language modeling, acoustic models, speech comprehension, and dialogue estimation. Although traditional GMMs—HMM-based machine learning models—seem better suited to speech spectrogram features, recent experimental results have clearly demonstrated that speech spectrogram features are more deeply learned with neural network-based machine learning models such as GMMs and deep neural networks than with HMMs [4]. This study also highlights that performance enhancements may be obtained by fine-tuning the design of deep neural networks that are state-of-the-art in both computational and phonological terms for automated spoken language recognition. Over the past decade, great progress has been made in the field of spoken language recognition owing to the current advancements in signal processing and cognitive research [7]. Several critical issues in language recognition, such as language classification, modelling methodologies, and software development strategies, have been addressed. The findings show that even though this part of the country has vastly grown in the last several years, it is far from being completely developed, notably in terms of linguistic characterization. Furthermore, this article offers an overview of the current research trends and future objectives, as well as the research techniques and technologies used in the development of the NIST-developed language recognition evaluation (LRE). An exhaustive survey of contemporary noise-robust speech recognition algorithms has been conducted over the past three decades [9]. Greater focus was placed on the established approaches, which are expected to retain and increase their usefulness in the future. The methods under examination were examined and their characteristics were assessed using five different metrics: using prior knowledge about acoustic environment distortion, domain processing method (e.g., model processing versus feature processing) versus process method (e.g., uncertainty processing versus predefined processing), using environment distortion models, and finally using trained acoustic models from the same adaptation process utilized in the testing stage[5]. This research offers information about resilient noise approaches and differentiates among various strategies, making it useful for readers.

3 Methodology

1. **Deep Learning Models**

One of the deep learning models often used for image categorization is the convolutional neural network (CNN), sometimes known as convent or CNN. Convolution, pooling, and fully linked layers are all key operations in a CNN. Several studies have also included a batch normalization layer. Originally developed for image classification, CNN may be used for sequential data by employing 1-dimensional convolutions instead of traditional 2-dimensional convolutions. Speaker identification systems include 1-dimensional CNN. A similarity function can be trained by employing a Siamese network that consists of two identical networks sharing weights. In [12], a one-shot learning assignment for face verification was performed using a Siamese network. A Siamese network typically employs two identical CNNs to learn the feature representation and compares the similarities between the two inputs. Because the main goal is to determine whether two input voices belong to the same person, this problem can be approached as binary classification. Another frequent way to learn from sequential data is through a long short-term memory network, also known as an LSTM network. Long-term dependencies in the original recurrent neural network were a problem that [10] sought to solve with the LSTM algorithm. The input, forget, candidate, and output gates comprise LSTM, which controls the learning process via four gates. To make things even more complicated, each LSTM cell has two distinct states: visible (cell state) and concealed (hidden). From one moment in time to the next. Both [11], where the authors used LSTM to solve large-vocabulary speech recognition problems, and [8], which uses LSTM for both voice enhancement and automatic speech recognition, have also looked at LSTM in depth.

2. **Speech Recognition**

Speech recognition is described as the process of voice identification based on spoken words by converting a signal acquired by audio equipment in a book on artificial intelligence (AI) [13]. Speech recognition is also a mechanism for recognizing human speech commands and converting them into data that computers can access. It is important to understand the difference between speech and sound because speech has different signal characteristics. Making a sound recognizable or identifiable so that it can be used necessitates many efforts, including voice or speech recognition. Acoustic-phonetic voice recognition, artificial intelligence voice recognition, and pattern recognition are three approaches to voice recognition. Figure 1 shows a block diagram that explains the pattern recognition approach to speech recognition Fig. 1 [14].

3. **Convolutional Neural Networks**

Deep learning algorithms such as convolutional neural networks (CNNs) are commonly used in ASR systems. Weight sharing, convolutional filters, and pooling are just a

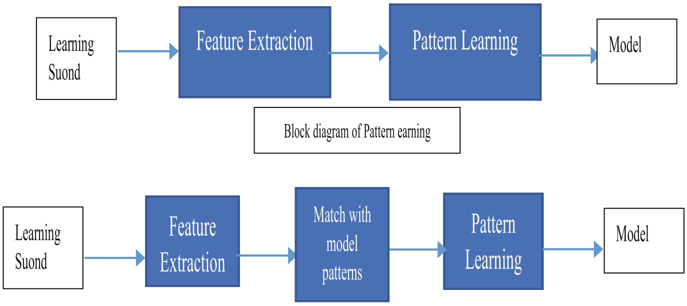


Fig. 1. Speech recognition block diagram

few of the appealing developments in CNNs that have been made. Consequently, CNN's outperformed other ASR technologies in this area. Multiple convolutional layers are used in CNNs to achieve complex functionality. A block diagram of the CNN is shown in Fig. 2.

4. Automatic Speech Recognition

Figure 3 depicts the architecture of an ASR system that uses ASR. This model comprises four parts: signal processing and feature extraction, an acoustic model (AM), a linguistic model (LM), and a hypothetical search. As an audio signal is sent to the system, it is processed to remove noise and channel distortion and then converted to the frequency domain, where it may be extracted as a vector feature that stands out.

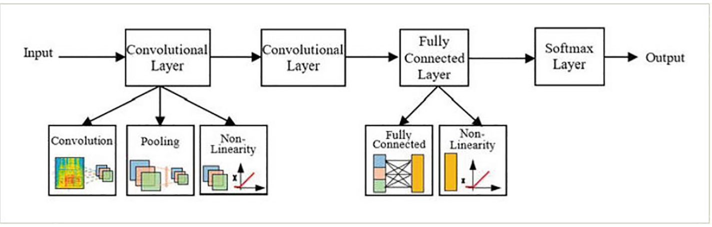


Fig. 2. Block diagram of CNN.

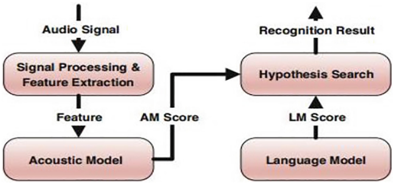


Fig. 3. Architecture of an ASR System

5. **Python**

Python is an object-oriented and highly structured programming language with dynamic semantics. Scripting or glue language with a higher level of knowledge structure mixed with dynamic and dynamic binding makes it highly marketable for quick creation, as well as the use of existing components. Python's easy-to-read syntax promotes readability while lowering software maintenance costs. Because many offices provide information analysis and data classification, Python supports work on all artificial intelligence algorithms, especially in the field of machine learning and deep learning, which is the core of our [15], as it provides many offices with small code and high and accurate performance in the field of information analysis and data classification.

6. **PyCharm**

Python programming was made easier with PyCharm, an integrated development environment (IDE). A JetBrains subsidiary in the Czech Republic created it [16]. In addition to code analysis, it has a graphical debugger, an integrated unit tester, and support for version control systems (VCSes). [15] PyCharm is available for Windows, macOS, and Linux and is, therefore, a cross-platform.

7. **Python GUI-Tkinter**

Python has several graphical user interfaces (GUI) creation options. Tkinters are the most widely used graphical user interface method. The Tk GUI toolbox in Python has a normal Python interface. The fastest and easiest way to create GUI applications is to use Python with a Tkinter. Creating a graphical user interface with a Tkinter is an easy task.

8. **Google Cloud Speech**

To exploit Google's speech recognition capabilities, a Google Cloud Speech API was created. This API exhibits an excellent speech recognition performance. A neural network was used to recognize 120 languages and dialects. It is possible to use real-time streaming or file input. Because it is a Web API, you will need a way to connect to the internet.

4 Design

The application is designed using the Python language, which supports an interactive user interface that helps users use the application easily:

5 Implementation

In this case, speech recognition systems can be classified by their elegance level.

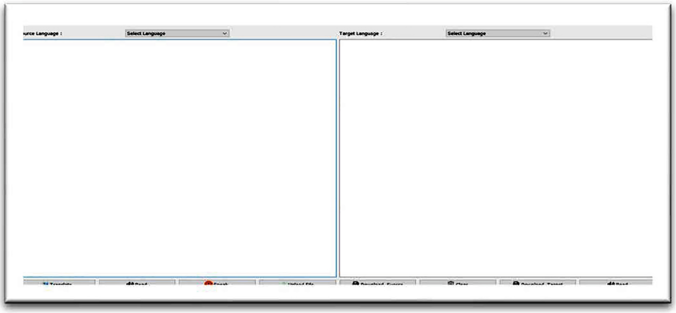


Fig. 4. Snapshot of Main Interface

5.1 Type of Speech

5.1.1 Isolated Word

An isolated word that understands the concept of "attain" usually means that the following two conditions must be met: quietness on both sides of the test window. This method is suitable for single or short utterances. This is a state that has "Listen and Not Listen" written all over it. Another name for this class may be “isolated utterance.“ [6].

5.1.2 Connected Word

Similar to isolated words, connected word systems are similar to words that are not linked together yet allow independent sentences to be uttered with a minimal pause between them.

5.1.3 Continuous Speech

Artificial intelligence programs that provide continuous voice recognition are said to enable users to talk nearly naturally while the machine decides on the content[1]. Some of the most challenging to develop are continuous speech recognizers because they use unique techniques to identify speech segments.

5.1.4 Spontaneous Speech

This may be defined as natural-sounding speech that has not been rehearsed. If a computer has an ASR system with the capacity to process spontaneous speech, it will be able to handle a wide range of natural speech characteristics, such as running words together.

5.2 Application Work

5.2.1 Run Application

This is the main interface of the application when it starts running (Fig. 5).

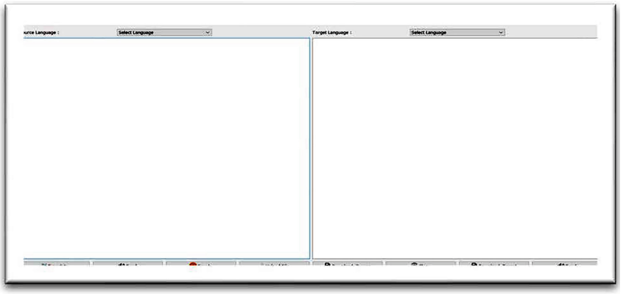


Fig. 5. Snapshot of Main Interface.

5.2.2 Select Source Language

Users can select any source language.

E.g: We will select the English language (Fig. 6).

5.2.3 Select Target Language

Users can select any target language.

E.g: We will select the Arabic language (Fig. 7).

5.2.4 Speak via Microphone

Users can press the speak button to start speaking in lectures.

E.g: We will say "hello how are you" (Fig. 8).

In Fig. 4, the user presses the speaking button to start speaking, and the application listens to and recognizes the sound of the user to convert speech to text and translate it.

5.2.5 Download Source Speech and Target After Translate

Users can download source text as a pdf file or print it using a printer (Fig. 9).

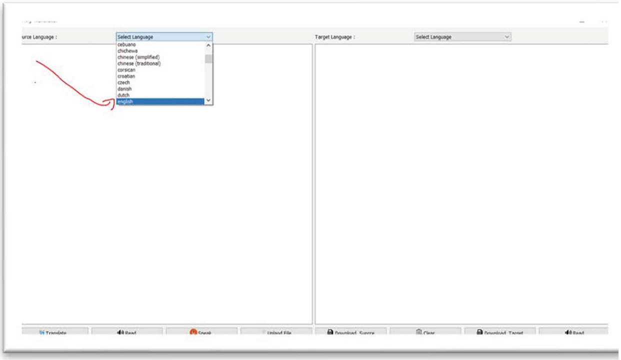
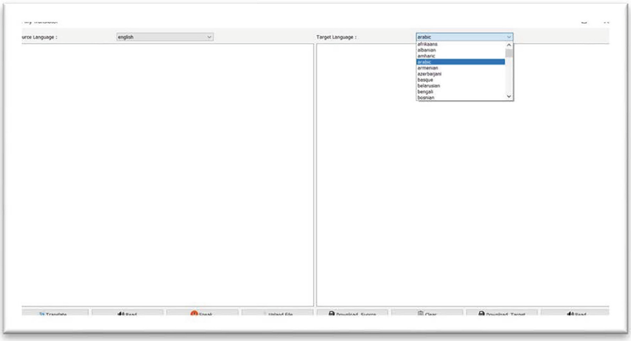
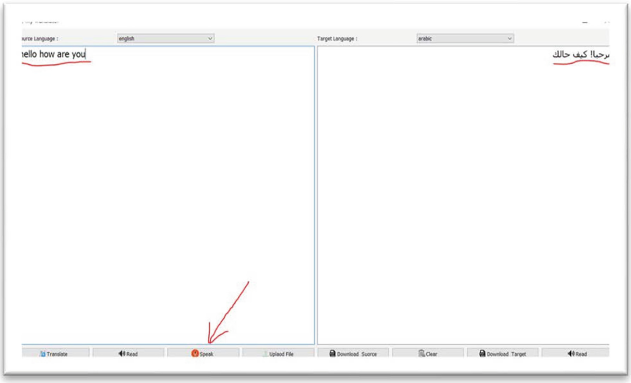


Fig. 6. Snapshot of Select Source language.


urce Language : english

Target Language : arabic

Fig. 7. Snapshot of Select target language

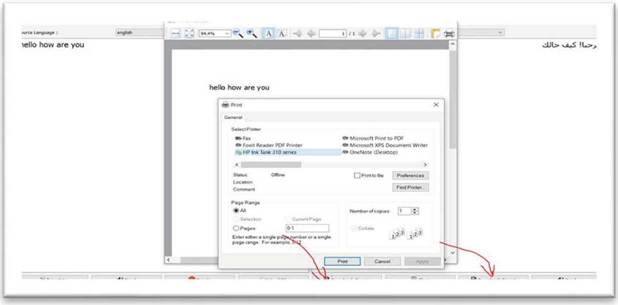

urce Language : english

Target Language : arabic

ıello how are you

رحبا! كيف حالك

Fig. 8. Snapshot of Speak Process.


urce Language : english

ıello how are you

رحبا! كيف حالك

94.4%

1 / 1

hello how are you

Print

General

Select Printer

- Fax
- Foxit Reader PDF Printer
- HP Ink Tank 310 series
- Microsoft Print to PDF
- Microsoft XPS Document Writer
- OneNote (Desktop)

Status: Offline

Location:

Comment:

□ Print to file

Page Range

● All

○ Selection

○ Current Page

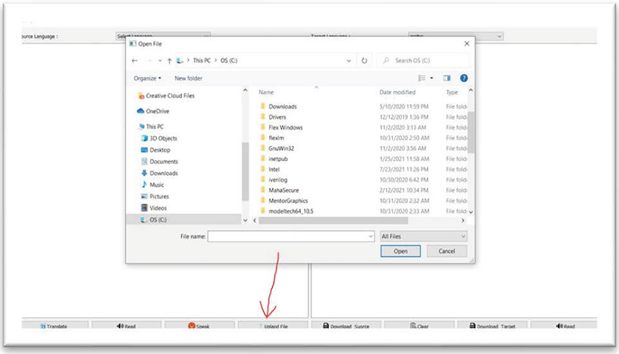
○ Pages: 0-1

Enter either a single page number or a single page range. For example, 5-12

Number of copies: 1

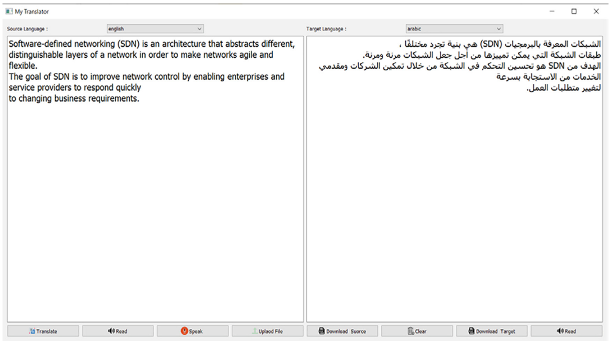
☑ Collate

Fig. 9. Snapshot of downloading source text and sending it to printer



Name	Date modified	Type
Downloads	5/10/2020 11:59 PM	File fold
Drivers	12/12/2019 1:36 PM	File fold
Flex Windows	11/2/2020 3:13 AM	File fold
flexlm	10/31/2020 2:50 AM	File fold
GnuWin32	11/2/2020 3:56 AM	File fold
inetpub	1/25/2021 11:58 AM	File fold
Intel	7/23/2021 11:26 PM	File fold
iverilog	10/30/2020 6:42 PM	File fold
MahaSecure	2/12/2021 10:34 PM	File fold
MentorGraphics	10/31/2020 2:32 AM	File fold
modeltech64_10.5	10/31/2020 2:33 AM	File fold

Fig. 10. Snapshot of Upload file



Source Language : english

Target Language : arabic

Software-defined networking (SDN) is an architecture that abstracts different,
distinguishable layers of a network in order to make networks agile and
flexible.
The goal of SDN is to improve network control by enabling enterprises and
service providers to respond quickly
to changing business requirements.

الشبكات المعرفة بالبرمجيات (SDN) هي بنية تجرد مختلفًا ،
طبقات الشبكة التي يمكن تمييزها من أجل جعل الشبكات مرنة ومرنة.
الهدف من SDN هو تحسين التحكم في الشبكة من خلال تمكين الشركات ومقدمي
الخدمات من الاستجابة بسرعة
لتغيير متطلبات العمل.

Fig. 11. Snapshot of after Upload file

5.2.6 Upload File

Users can upload the files to translate to any language (Figs. 10 and 11).

6 A Multilingual Voice Translation Processing Architecture That Handles Multilingual Speech

In Fig. 12, an English phrase is transformed into text and subsequenty translated into Arabic, which is then multilingual speech recognition module examines the input speech, comparing it to a large amount of speech data, which are created by representing all the phonemes found in each speech utterance in the English syllabify. Next, the string of phonemes was transformed into a string of words written in the English writing system, resulting in a string of words having the greatest possible probability. An English sentence is created using an engine trained on large quantities of English text by examining the probability of the occurrence of a string of three words. Each English word in the string was replaced with an appropriate Arabic word using a conversational-language translation module. To provide a sequence, the order of Arabic words is altered [7]. A

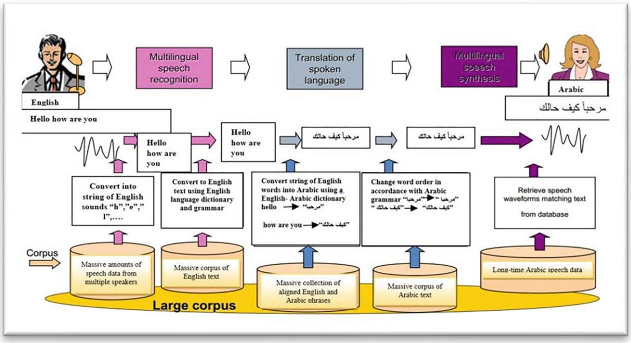

Multilingual speech recognition
Translation of spoken language
Multilingual speech synthesis
English
Hello how are you
Arabic
مرحباً كيف حالك
Hello how are you
Hello how are you
مرحباً كيف حالك
مرحباً كيف حالك
Convert into string of English sounds "h","e"," l",....
Convert to English text using English language dictionary and grammar
Convert string of English words into Arabic using a English- Arabic dictionary hello → "مرحبا"
how are you → "كيف حالك"
Change word order in accordance with Arabic grammar "مرحبا" → "مرحبا"
"كيف حالك" → "كيف حالك"
Retrieve speech waveforms matching text from database
Corpus
Massive amounts of speech data from multiple speakers
Massive corpus of English text
Massive collection of aligned English and Arabic phrases
Massive corpus of Arabic text
Long-time Arabic speech data
Large corpus

Fig. 12. The Architecture of the speech translation system

translated English-to-Arabic translation model was used to replace the original English words in the strings. An engine trained on large quantities of Arabic text generates an Arabic string of words with the correct probability of occurrence to rearrange them. Once this has been sent to the speech synthesis module, it is sent to the end-user. Speech synthesis connects English words to associated speech sounds, calculates pronunciation and intonation, and selects sounds from a long-term speech data database. Using statistical modeling and machine learning, the method of speech recognition and synthesis known as "corpus-based speech recognition and synthesis" applies massive speech corpora.

7 Discussion

In this study, we focus on how the application fulfills universities' needs for communication improvement between lecturers and students. Artificial intelligence techniques were used to recognize speech, translate speech, convert speech into text, and then save the lecture into text and audio files.

- The lecturers must speak clearly.
- The lecturers should use a wireless microphone that is connected to a computer.
- The application will install on the classroom computer.
- A user interface was designed to facilitate the selection of the present language and target language.
- Creating a display screen for students that displays the process of converting speech to text at present and translating it into the target languages.
- The speech that has been converted into text depends on the translation process; thus, by pressing a translation button, the entire speech is translated.
- At the end of the lecture, the student will be able to save and download the lecture as text, pdf, and audio files by pressing a button.
- In addition, the user can upload a text file to translate it into any language, whether to text or audio files.

8 Conclusion

In this paper, we have included many papers that helped us develop the aforementioned application, as stated above. This application was developed using a high-level programming language called Python. There is a possibility that the application might change educational access for the disabled and for all people. The project has already shown significant gains in the overall understanding and knowledge of how voice recognition, speech-to-text, and translation might be used in educational settings. Project success has a positive effect on business sector support and the consortium of universities that participate in the project. The project's programmer believes that it will attract a large audience due to the emphasis on student accommodation in classrooms. Our view is that a project's objective is to provide everyone with equitable access to knowledge.

References

1. Gaikwad S.K., Gawali B.W., and Yannawar, P., 2010. A review on speech recognition technique. International Journal of Computer Applications, 10(3), pp.16-24.
2. Nimbhore S. , Ramteke G., Ramteke R. ," Pitch Estimation of Marathi Spoken Numbers in Various Speech Signals", International Conference on Communication and Signal Processing, April 3–5, 2013.
3. More S. , P. Borde, S Nimbhore," Isolated Pali Word (IPW) Feature Extraction using MFCC & KNN Based on ASR", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278–0661, p-ISSN: 2278–8727, Volume 20, Issue 6, Ver. II, Nov - Dec 2018.
4. Nimbhore S., Mache S., "Processing of Devnagari Text to Speech Synthesis: A Review, International Journal of Management, Technology And Engineering Volume IX, Issue I, JANUARY/2019 ISSN NO: 2249–745.
5. Morgan N., 2011. Deep and wide: Multiple layers in automatic speech recognition. Ieee transactions on audio, speech, and language processing, 20(1), pp.7-13.
6. Deng, L., Li J., Huang J.T., Yao K., Yu D., Seide F., Seltzer M., Zweig G., He X., Williams J. and Gong Y., 2013, May. Recent advances in deep learning for speech research at Microsoft. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 8604–8608). IEEE.
7. Li H., Ma B. and Lee K.A., 2013. Spoken language recognition: from fundamentals to practice. Proceedings of the IEEE, 101(5), pp.1136-1159.
8. Chen Z., Watanabe S., Erdogan H., and Hershey J. R.2015. Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks. In Proceedings of the 16th Annual Conference of the International Speech Communication Association (Dresden, Germany, September 6–10, 2015). INTERSPEECH '15. 3274–32780.
9. Chowdhury A. and Ross A. 2017. Extracting sub-glottal and supra-glottal features from MFCC using convolutional neural networks for speaker identification in degraded audio signals.In Proceedings of the 2017 IEEE International Joint Conference on Biometrics (Denver, CO, United States, October 1–4, 2017). IJCB '17. IEEE, New York, NY, 608–617.OI=https://doi.org/10.1109/BTAS.2017.8272748.
10. Hochreiter S. and Schmidhuber J. 1997. Long short-term memory. Neural Computation 9, 8 (Nov. 1997), 1735–1780.DOI=https://doi.org/10.1162/neco.1997.9.8.1735.
11. Li X. and Wu X. 2015. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (Brisbane, Australia, April 19–24,

2015). ICASSP'15.IEEE,NewYork, NY, 4520–4524.DOI=https://doi.org/10.1109/ICASSP.2015.7178826.

12. Taigman Y., Yang M., Ranzato M., and Wolf L. 2014.Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Columbus, OH, United States, June 23–28, 2014). CVPR '14. IEEE, New York, NY, 1701–1708. https://doi.org/10.1109/FCVPR.2014.220.
13. Amrizal V., Q Aini. Artificial Intellegence (in Indonesia Kecerdasan Buatan). Jakarta Barat. Halaman Moeka Publishing. 2013. E Widiyanto, SN Endah, S Adhy. Speech Application to Text in Bahasa using Mel Frequency Cepstral Coefficients and Hidden Markov Models (in Indonesia Aplikasi Speech To Text Berbahasa Indonesia Menggunakan Mel Frequency Cepstral Coefficients Dan Hidden Markov Model). Prosiding Seminar Nasional Ilmu Komputer Undip. 2014: 39-44.
14. Endah SN., Adhy S., Sutikno S. Comparison of Feature Extraction Mel Frequency Cepstral Coefficients and Linear Predictive Coding in Automatic Speech Recognition for Indonesian.TELKOMNIKA Telecommunication Computer Electronics and Control. 2017; 15(1): 292.
15. AL-Bakhrani Ali A., et al. "Machine Learning and Deep Learning to Do Early Predictions of COVID-19 Infection Using Chest X-Ray Images." Machine Learning 62.07 (2020).
16. Matthew D.Z., Rob F., Visualizing and Understanding Convolutional Networks, Springer International Publishing Switzerland 2014, pp. 818–833.